Lecture Notes on Data Engineering and Communications Technologies 139

Vijayan K. Asari
Vijendra Singh
Rajkumar Rajasekaran
R. B. Patel *Editors*

Computational Methods and Data Engineering

Proceedings of ICCMDE 2021

Springer

Lecture Notes on Data Engineering and Communications Technologies

Volume 139

The aim of the book series is to present cutting edge engineering approaches to data technologies and communications. It will publish latest advances on the engineering task of building and deploying distributed, scalable and reliable data infrastructures and communication systems.

The series will have a prominent applied focus on data technologies and communications with aim to promote the bridging from fundamental research on data science and networking to data engineering and communications that lead to industry products, business knowledge and standardisation.

Indexed by SCOPUS, INSPEC, EI Compendex.

All books published in the series are submitted for consideration in Web of Science.

Vijayan K. Asari · Vijendra Singh ·
Rajkumar Rajasekaran · R. B. Patel
Editors

Computational Methods and Data Engineering

Proceedings of ICCMDE 2021

Springer

Editors
Vijayan K. Asari
Department of Electrical and Computer Engineering
University of Dayton
Dayton, OH, USA

Vijendra Singh
School of Computer Science
University of Petroleum and Energy Studies
Dehradun, Uttarakhand, India

Rajkumar Rajasekaran
School of Computer Science and Engineering
Vellore Institute of Technology
Vellore, Tamil Nadu, India

R. B. Patel
Department of Computer Science and Engineering
Chandigarh College of Engineering and Technology
Chandigarh, India

ISSN 2367-4512 ISSN 2367-4520 (electronic)
Lecture Notes on Data Engineering and Communications Technologies
ISBN 978-981-19-3014-0 ISBN 978-981-19-3015-7 (eBook)
https://doi.org/10.1007/978-981-19-3015-7

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

We are pleased to present Springer Book entitled *Computational Methods and Data Engineering*, which consists of the Proceedings of the International Conference on Computational Methods and Data Engineering (ICMDE 2021) papers.

The main aim of the International Conference on Computational Methods and Data Engineering (ICMDE 2021) was to provide a platform for researchers and academia in the areas of computational methods and data engineering to exchange research ideas and results and collaborate together. The conference was held at the School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India, during November 25–26, 2021.

All the 41 published chapters in the Computational Methods and Data Engineering book have been peer-reviewed by the three reviewers drawn from the scientific committee, external reviewers, and editorial board depending on the subject matter of the paper. After the rigorous peer review process, the submitted papers were selected based on originality, significance, and clarity and published as book chapters.

We would like to express our gratitude to the management, faculty members, and other staff of the School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India, for their kind support during the organization of this event. We would like to thank all the authors, presenters, and delegates for their valuable contribution to making this an extraordinary event.

We would like to acknowledge all the members of honorary advisory chairs, international/national advisory committee members, general chairs, program chairs, organization committee members, keynote speakers, the members of the technical committees, and reviewers for their work.

Finally, we thank series editors, Advances in Intelligent Systems and Computing, Aninda Bose, Sharmila Mary Panner Selvam, and Radhakrishnan for their high support and help.

Dehradun, India — Prof. Vijendra Singh
Vellore, India — Prof. Rajkumar Rajasekaran
Dayton, USA — Prof. Vijayan K. Asari
Chandigarh, India — Prof. R. B. Patel

Contents

About the Editors

Dr. Vijayan K. Asari is Professor in Electrical and Computer Engineering and Ohio Research Scholars Endowed Chair in Wide Area Surveillance at University of Dayton, Dayton, Ohio. He is Director of the University of Dayton Vision Lab (Center of Excellence for Computer Vision and Wide Area Surveillance Research). Dr. Asari had been Professor in Electrical and Computer Engineering at Old Dominion University, Norfolk, Virginia, till January 2010. He was Founding Director of the Computational Intelligence and Machine Vision Laboratory (ODU Vision Lab) at ODU. Dr. Asari received the Bachelor's degree in electronics and communication engineering from the University of Kerala (College of Engineering, Trivandrum), India, in 1978, the M.Tech. and Ph.D. degrees in Electrical Engineering from the Indian Institute of Technology, Madras, in 1984 and 1994, respectively.

Dr. Vijendra Singh is working as Professor in School of Computer Science at The University of Petroleum and Energy Studies, Dehradun, Uttarakhand, India. Dr. Singh received his Ph.D. degree in Engineering and M. Tech. degree in Computer Science and Engineering from Birla Institute of Technology, Mesra, India. He has 20 years of experience in research and teaching including IT industry. Dr. Singh's major research concentration has been in the areas of data mining, image processing, Big data, machine learning, and Deep Learning. He has published more than 85 scientific papers in this domain. He has served as Editor in Chief, *Procedia Computer Science*, Vol 167, 2020, Elsevier; Editor in Chief, *Procedia Computer Science*, Vol 132, 2018, Elsevier; Associate Editor, *BMC Medical Informatics and Decision Making*, Springer Nature; *International Journal of Healthcare Information Systems and Informatics*, IGI Global, USA; Editor, *BMC Digital Health*, Springer Nature; Editorial Board Member, *International Journal of Information and Decision Sciences*, Inderscience, UK. He received the IBM Edu Leader Award. Vijendra Singh is Member of ISTE, IEEE, and ACM.

Dr. Rajkumar Rajasekaran is Professor and Researcher, at the Vellore Institute of Technology, India. He earned his Ph.D. in Computer Science and Engineering. He is Professor of Data Analytics at School of Computer Science and Engineering at VIT University in Vellore, India. In his current position, he is responsible for teaching data science, data visualization, and big data analytics. He was Head of data analytics research group. Research focuses on healthcare data analytics, agricultural data analytics, and streaming data analytics. He is Specialist in exploratory data analytics, data visualization, and the development and applications of agricultural data analytics techniques particularly implementing with R programming. He has published more than 60 papers in Scopus and SCI and chaired many conferences and published three authored books. He has got 6 Indian patents and 1 Australian patent. His current research looks at the healthcare data analytics of international governance by emerging countries like India. Rajkumar has learned and applied data analytics in healthcare and agriculture data combining case study analysis.

Prof. R. B. Patel is working as Professor in the Department of Computer Science and Engineering, Chandigarh College of Engineering and Technology (CCET), Chandigarh, India. Prior to joining the CCET, he worked as Professor at NIT, Uttarakhand, India, and Dean, Faculty of Information Technology and Computer Science, Deenbandhu Chhotu Ram University of Science and Technology, Murthal, India. His research areas include mobile and distributed computing, machine and deep learning, and wireless sensor networks. Prof. Patel has published more than 150 papers in international journals and conference proceedings. He has supervised 16 Ph.D. scholars and currently 02 are in progress.

A Graph-Based Extractive Assamese Text Summarization

Nomi Baruah, Shikhar Kr. Sarma, Surajit Borkotokey, Randeep Borah, Rakhee D. Phukan, and Arjun Gogoi

Abstract This paper proposes an approach to introduce a text summarization in the Assamese language as it has been observed that text summarization is an important natural language processing application developed to condense the original or main text by retrieving the relevant information while retaining the main matter of the text. In the proposed approach, we have presented our idea of text summarization with the help of a graph-based system where we have denoted the document as a graph and in which the vertices of the graph depict the sentences. For our objective, Assamese WordNet has been considered in the whole text summarization process and the said process includes three major phrases: Pre-processing phrase, features extraction and graph construction stage with finally applying LexRank and extraction of summary. We believe that this proposed approach yields better results when contrasted with the statistical approach and previous graph-based approaches in Assamese in terms of usage of nouns and amalgamation of semantic similarity and statistical features.

Keywords Assamese · Cosine similarity · Extractive · Statistical features · WordNet

1 Introduction

The Internet, daily basis, is flooded with an immense volume of content or data online, and it is felt that extracting the relevant information from the targeted text is tedious and cumbersome. Thus, the researchers have come up with a solution and proposed a text summarization approach that curtails and extracts the relevant words from the input text and generates a text in the form of a summary while preserving the meaning of the input text and help readers or users to get the desired data without

N. Baruah (✉) · S. Borkotokey · R. Borah · R. D. Phukan · A. Gogoi
Dibrugarh University, Dibrugarh 786004, India
e-mail: baruahnomi@gmail.com

S. Kr. Sarma
Gauhati University, Guwahati, India

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_1

then need to go through the whole text [1]. Text summarization can be defined as an automated process of truncating an amount of text from the original or input data and extract the most relevant text to generate a summary for its users. While generating the summary, it is also kept in mind that the yielded content holds up its original meaning. In terms of implementation of text summarization in different languages, it is observed that the number of researches conducted is more for the English language compared to other languages because of its structure, grammar and less or minimum morphological intricacies, whereas, the Assamese language is characterized by a complex structure and morphology. And considering the usage of the Assamese language, i.e. 15.4 million people speak this language across the globe [2], the necessity of Assamese text summarization is felt.

Depending on the factor, text summarization is classified into multiple categories for comparison purposes. Moreover, based on the number of documents, the text summarization approach is divided a single document and multiple document summarization. In this approach, considering the nature of retrieval of sentences, it is categorized as the extractive summary and abstractive summarization. An extractive summary is defined as the process in which the sentences are extracted from the original input text without making any kind of modifications to the sentence structure. The abstractive summary is referred to the approach that generates a summary without adhering to the sentence structure. The text summarization approach can further be divided into general and query-based summaries. The general summary generates sentences without having any relation to the title or any question, whereas the query-based summary is returned because of the correlation between the summary a query submitted or the relationship between the document title and sentence [3].

The headway in Assamese text summarization has been gradual because it suffers from the lack of researches in natural language processing (NLP). It may also be noted that nouns, a part of speech, have a direct impact on the relevancy of the sentence and with the inclusion of more nouns the relevance of the sentence enhances. Thus, this proposed paper is an attempt to explore a novel approach that depends on the amalgamation of graph theory and nouns count in sentences to draw out Assamese text summarization. The approach initiates with the implementation of the following pre-processing techniques: normalization, tokenization, removal of stop words and stemming. The requisite features are extracted and a graph is built, and LexRank algorithm is applied to extract the summary and filter redundancy. In this process, the sentences having senses are sorted by computing the final score, and then the summary is extracted and inessential sentences are eliminated. Further works will be discussed in Sect. 2. Section 3 elaborates the motivation and problem statement in the proposed approach. Section 4 discusses the proposed approach, Sect. 5 discusses the results and Sect. 6 about the conclusion, respectively.

This paper primarily presents a graph-based text summarization approach for a single Assamese document by considering the sentence extraction methodology. In this proposed approach, the statistical features and semantic similarity is consolidated to look over their effect on the quality of the extracted summary of the Assamese text. Cosine similarity is also applied in this approach to recognize the optimal sentence considered for a combination of summary based on maximizing informative sores and cohesion between sentences. Cohesion is the scope to which the ideas in the text are communicated vividly and in a systematic fashion of one another. The objective of this research can be developed into two research questions. (1) Does the usage of 'noun' words enhance the performance of the summary? (2) Does the amalgamation of semantic similarity and statistical features generate a better summary in terms of sense regarding ROUGE scores for Assamese text?

2 Related Work

Several types of research were carried out in Assamese text summarization and the same are discussed as follows.

Kalita et al. [4] propounded an approach on automatic text summarization while emphasizing Assamese WordNet and considering a stop word list. The considered WordNet database was diminutive and so the author augmented words of text document from his part. For the said purpose, the author experimented on ten documents.

Baruah et al. [5] expounded an approach using statistical features and Assamese WordNet. The varied features considered to identify the salient sentences to produce an effective summary aspect such as TF-IDF, sentence position, sentence length and numerical identification are examined. The efficacy of the approach proposed by the author is exhibited through a set of experiments considering ROUGE measure, and the evaluation is represented in terms of precision, recall and F-1 score.

Baruah et al. [6] put forward an approach by applying statistical features to extract a summary while assimilating statistical and linguistic features. The approach is also an attempt to extract the salient points from the input Assamese document and includes it in the summary. Additionally, the statistical features considered in this approach applied a new modified tf-idf and other features such as sentence length, sentence positioning and recognition of numerical. Apart from the cited statistical features, the linguistic features are also considered. The linguistic features applied to identify the cue words, cue phrases and nouns. In this approach, summaries are extracted both automatically and by humans. And all the summaries extracted are contrasted by considering the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) framework.

Baruah et al. [7] proposed an approach by applying statistical features, graph-based features and Assamese WordNet. The statistical features applied are to be cited as TF-IDF and sentence position to evaluate the sentences of the input document. The evaluation of the sentences is done based on the assimilation of semantic and statistical traits and considering the significance of sentences and diversity. Further, synsets from the WordNet to compute the word frequency. Apart from recognizing and weighing the salient sentences, the formula of cosine similarity and vector centrality is used, and the result derived is illustrated in graph representation.

3 Motivation and Problem Statement

It is to be mentioned that very little researches are conducted on the Assamese summarization system, and observations of the researches have revealed that nouns enhance the significance of the sentences. In other words, it may be implying that more nouns carry more information in the sentence. A few types of research were also conducted in Assamese text summarization using the graph-based method and a modified TF-IDF and a few nouns in the sentences to calculate the initial rank of the targeted sentences. Apart from a few steps, the text summarization approach in Assamese is similar to English. For instance, an algorithm may be developed for text summarization in English to identify the words with upper-case or lower-case letters, whereas no requirement arises for Assamese. The proposed research in this paper also considered stemming to stem the words to its root word. The stemming concept considered is mentioned in [6].

4 Proposed Approach

In this section, the proposed approach is explained. Figure 1 illustrates the pictorial diagram of the proposed approach which has three major phrases. The first phase includes the pre-processing tasks, viz. normalization, tokenization, stopwords removal and stemming. In the second phase, the desired linguistic and statistical features are extracted. In the third phase, the ranking method is applied, and a summary is extracted, and following it, the performance is evaluated.

4.1 Stages of Proposed Approach

Stage 1: In this stage, the targeted document is entered and prepared for features extraction. Extraction of text is intended from a single document in Assamese by encoding UTF-8. For this objective, the input document is classified into paragraphs, and it is further divided into sentences and finally into words. In this process, an attempt is made to remove the stopwords to make the words more relevant for the intended purpose. Stemming is another step consider red for extraction of the root of every word in the sentence. It may kindly be noted that the process used for stemming is defined in [6].

Stage 2: At the second stage, calculation of the statistical features is done. Moreover, the significance of a word in a document is to be measured, and thus, it is done with the help of word frequency. The author considered the Assamese WordNet as described in [8], and the same is applied, and also, the word frequency is calculated again. It has been observed that the position of a sentence in text summarization plays a vital role, and hence, it is presumed that sentences at the beginning of a

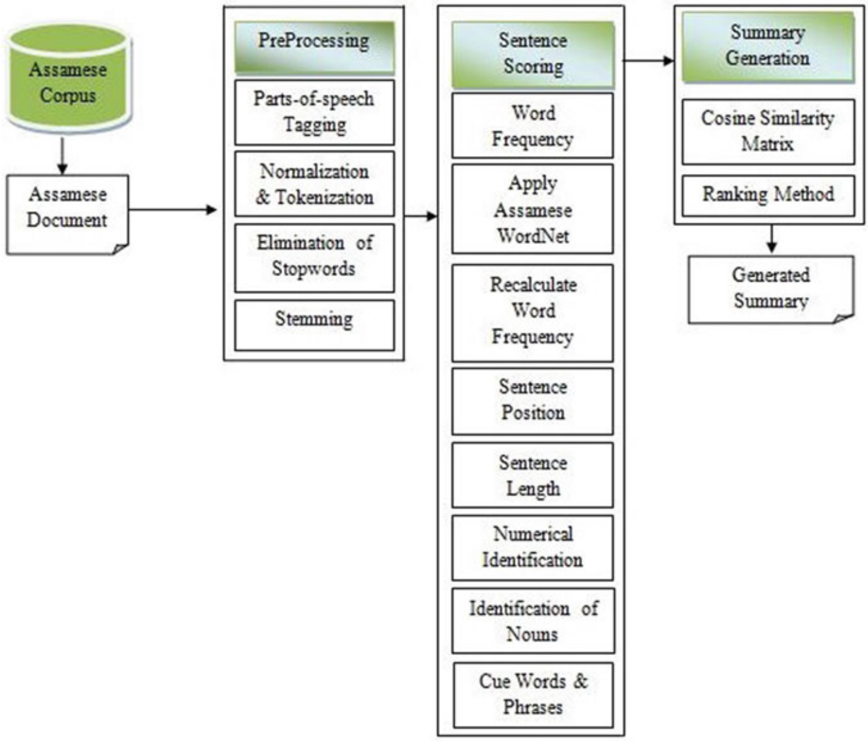


Fig. 1 Proposed text summarization approach

paragraph contain more information in comparison to the ones that are used towards the end of the text as defined by [9]. Length of sentences is another feature that is taken into account in document summarization. To evaluate the importance of a sentence, each sentence is assigned a score while considering its length and is usually carried out subject to the condition of the existence of words in the sentence post-pre-processing. It is also observed that a long sentence has more probability to contain relevant information compared to a short sentence. The length of the sentence is defined as [9].

In this proposed approach, the sentences that carry nouns are identified and assimilated in the summary [10]. In Assamese, 'proper noun' are words that indicate person, place, thing, animals, concept, etc. But it is to be noted that all proper nouns in Assamese may or may not have been added in the Assamese Dictionary. The cue words and phrases that are present in a document are incorporated in the summary, and the score of this feature computed is defined by [11]. The salient sentences are extracted from the document calculating the total score that is allotted to them in this process.

$$\mathrm{Stotal}_i = \text{Score of word frequency after application of Wordnet}$$

$$\begin{aligned} &+ \text{ score of sentence position} + \text{score of sentence length} \\ &+ \text{ score of Numerical Identification} \\ &+ \text{ score of Identification of nouns} \\ &+ \text{ score of Cue words and phrases} \end{aligned} \tag{1}$$

where Stotal_i refers to the total score of each sentence i from a document D.

The score of each sentence in the document is calculated using Eq. 1. Based on the comparison ratio, the top scored sentences are considered for the summary.

Stage 3: In this phase, informative abstracts are not only derived from statistical features as the meanings of the words are not considered. The features of a good summary are informativeness, cohesion and coherence, but in this experment, it has been observed that the summaries that are generated considering extractive summarization system defect in the features such as cohesiveness and coherence [12]. It can be concluded that to generate a good summary, the corelation of sentences is essential in an input document, and the same has to be discovered which has to be done by computing using a cosine similarity measure. The calculation of cosine similarity between sentences is defined in [13], and the matrix is illustrated in a graph termed as Markov chain and is represented in Figs. 2 and 3.

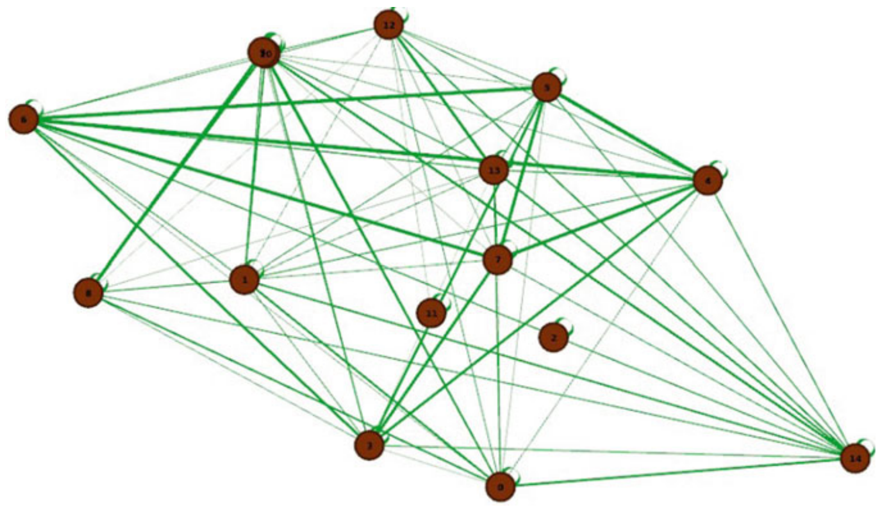
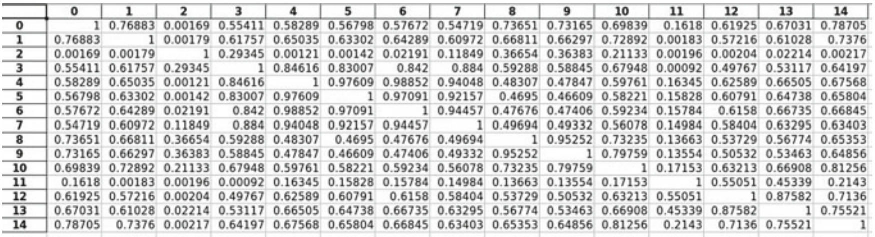


Fig. 2 Graph generated



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	1	0.76883	0.00169	0.55411	0.58289	0.56798	0.57672	0.54719	0.73651	0.73165	0.69839	0.1618	0.61925	0.67031	0.78705
1	0.76883	1	0.00179	0.61757	0.65035	0.63302	0.64289	0.60972	0.66811	0.66297	0.72892	0.00183	0.57216	0.61028	0.7376
2	0.00169	0.00179	1	0.29345	0.00121	0.00142	0.02191	0.11849	0.36654	0.36383	0.21133	0.00196	0.00204	0.02214	0.00217
3	0.55411	0.61757	0.29345	1	0.84616	0.83007	0.842	0.884	0.59288	0.58845	0.67948	0.00092	0.49767	0.53117	0.64197
4	0.58289	0.65035	0.00121	0.84616	1	0.97609	0.98852	0.94048	0.48307	0.47847	0.59761	0.16345	0.62589	0.66505	0.67568
5	0.56798	0.63302	0.00142	0.83007	0.97609	1	0.97091	0.92157	0.4695	0.46609	0.58221	0.15828	0.60791	0.64738	0.65804
6	0.57672	0.64289	0.02191	0.842	0.98852	0.97091	1	0.94457	0.47676	0.47406	0.59234	0.15784	0.6158	0.66735	0.66845
7	0.54719	0.60972	0.11849	0.884	0.94048	0.92157	0.94457	1	0.49694	0.49332	0.56078	0.14984	0.58404	0.63295	0.63403
8	0.73651	0.66811	0.36654	0.59288	0.48307	0.4695	0.47676	0.49694	1	0.95252	0.73235	0.13663	0.53729	0.56774	0.65353
9	0.73165	0.66297	0.36383	0.58845	0.47847	0.46609	0.47406	0.49332	0.95252	1	0.79759	0.13554	0.50532	0.53463	0.64856
10	0.69839	0.72892	0.21133	0.67948	0.59761	0.58221	0.59234	0.56078	0.73235	0.79759	1	0.17153	0.63213	0.66908	0.81256
11	0.1618	0.00183	0.00196	0.00092	0.16345	0.15828	0.15784	0.14984	0.13663	0.13554	0.17153	1	0.55051	0.45339	0.2143
12	0.61925	0.57216	0.00204	0.49767	0.62589	0.60791	0.6158	0.58404	0.53729	0.50532	0.63213	0.55051	1	0.87582	0.7136
13	0.67031	0.61028	0.02214	0.53117	0.66505	0.64738	0.66735	0.63295	0.56774	0.53463	0.66908	0.45339	0.87582	1	0.75521
14	0.78705	0.7376	0.00217	0.64197	0.67568	0.65804	0.66845	0.63403	0.65353	0.64856	0.81256	0.2143	0.7136	0.75521	1

Fig. 3 Cosine similarity matrix of the document

5 Experimentation and Results

This section furnishes the results derived in this experiment for the implemented statistical text summarization approach and graph-based text summarization approach of Assamese language. Initially, delineation of the dataset used for evaluation is made which is followed by the results derived in the automatic evaluation. The results are analysed to conclude.

5.1 Dataset

In this experiment, we have considered the dataset developed by [14] to evaluate the proposed summarization approaches. Interestingly, the Assamese corpus, a dataset of Assamese words, has more than 20 lakh words. Additionally, the data collection articles are categorized as culture, economy, local news, inter- national news, religion, sports, etc. are described in Table 1. The dataset that is considered in this proposed experiment is used to evaluate the approaches as proposed against human-generated summaries that consist of 30 randomly selected Assamese documents belonging to varied categories of the corpus [15]. In the case of the traditional summary system

Table 1 Dataset description as given in [14]

Category	Expected words per category
Newspaper	337,500
Magazine	299,500
Science	11,625
Arts	113,000
Short fiction	120,000
Criticism	52,500
Theatre	75,000
Novel	300,000
Trivia	15,000

generation approach, more than two model summaries are applied to evaluate as humans may disagree with each other while picking the summary of the document. Opinions of linguistics were also sought in this regard by requesting them to read and prepare summaries of the targeted articles while emphasizing the most salient sentences to create an extractive summary. The language experts were asked to select while ensuring the limitation of extracting not more than half of the sentences in the article. Finally, for every document, two different model summaries were extracted, and each one is prepared by a different person. The dataset is available in UTF-8 encoding.

5.2 Evaluation Using ROUGE

For evaluating the proposed paper, we have also applied Recall-Oriented Understudy for Gisting Evaluation (ROUGE) which is introduced by [16] as a standard metric. The intent of applying ROUGE is for the reason that it corresponds suitably with human evaluation if similar content is provided in text summarization [17]. This standard metric depends on N-Gram co-occurrence, and it contrasts the content of the targeted summary with other referred summaries and recognizes the number of N-Gram of words. The authors in this proposed experiment also took into account diverse types of N-Grams for calculation purposes such as unigrams (ROUGE-1), bigrams (ROUGE-2) and trigrams (ROUGE-3).

It may be noted that in both the summaries, i.e. automated and human-generated excessive unigrams, bigrams and trigrams are computed by applying ROUGE-1, ROUGE-2 and ROUGE-3, respectively. The results derived in this experiment are analysed to evaluate the performance of the proposed approaches. Moreover, the comparison is made between the proposed approaches to assess the performance of the approaches and is tested by using a 50 per cent compression ratio as described in [18].

Experiment 1: In this experiment, the final score of a sentence is calculated based on statistical features: TF-IDF, sentence position, sentence length, identification of numerical and cue words and phrases. The proposed summarizer automatically generates summaries of the ten documents from the corpus (Table 1).

Experiment 2: In this experiment, the final score of a sentence is calculated based on statistical features: TF-IDF, sentence position, sentence length, identification of numerical, identification of nouns and cue words and phrases. The proposed summarizer automatically generates summaries of the ten documents from the corpus (Table 2).

Experiment 3: In this experiment, the final score of a sentence is calculated based on statistical features: TF-IDF, sentence position, sentence length, identification of numerical, cue words and phrases and cosine similarity matrix. The proposed summarizer automatically generates summaries of the ten documents from the corpus (Table 3).

Table 2 ROUGE scores of statistical approach

Evaluation metric	ROUGE-1	ROUGE-2	ROUGE-3
Avg recall	0.5353	0.3630	0.3333
Avg precision	0.4950	0.3517	0.3237
Avg F-measure	0.5433	0.3877	0.3533

Table 3 ROUGE scores of statistical approach with identification of nouns

Evaluation metric	ROUGE-1	ROUGE-2	ROUGE-3
Avg recall	0.5662	0.3992	0.3766
Avg precision	0.5266	0.3880	0.3553
Avg F-measure	0.5660	0.4011	0.3838

Table 4 ROUGE scores of graph-based approach

Evaluation metric	ROUGE-1	ROUGE-2	ROUGE-3
Avg recall	0.6586	0.4789	0.4500
Avg precision	0.6324	0.4688	0.4343
Avg F-measure	0.6422	0.4778	0.4433

Table 5 ROUGE scores of graph-based approach with identification of nouns

Evaluation metric	ROUGE-1	ROUGE-2	ROUGE-3
Avg recall	0.6777	0.4950	0.4755
Avg precision	0.6672	0.4889	0.4011
Avg F-measure	0.6691	0.5021	0.4588

Experiment 4: In this experiment, the final score of a sentence is calculated based on statistical features: TF-IDF, sentence position, sentence length, identification of numerical, identification of nouns, cue words and phrases and cosine similarity matrix. The proposed summarizer automatically generates summaries of the ten documents from the corpus (Table 4 and 5).

Thus, it is evident from Tables 1 and 2 that the performance of the graph-based approach is superior to statistical approach, i.e. 21.36 percentage in terms of ROUGE-1,12.72 percentage in terms of ROUGE-2 and 12.63 percentage in terms of ROUGE-3.

5.3 *Comparison with Other Approaches*

Figure 1 and Fig. 2 represent the results obtained by comparing avg recall, avg precision and avg f-measure based on ROUGE-1 and ROUGE-2 where another approach is presented in [7]. The approach [7] is the latest graph-based text summarization approach in Assamese. It is seen that experiment 4 proved to generate quality-based summaries when compared with another graph-based approach. The summaries are compared with a 50% compression ratio.

In Figs. 4 and 5, the light blue colour refers to the implementation conducted using features in experiment 1. Brown colour refers to the implementation conducted using features in experiment 2. Light grey colour refers to the implementation

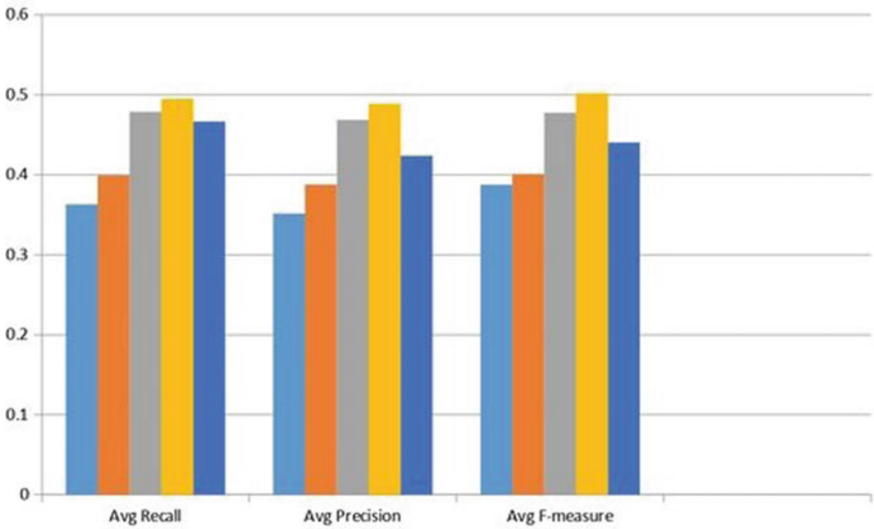


Fig. 4 ROUGE-1 scores

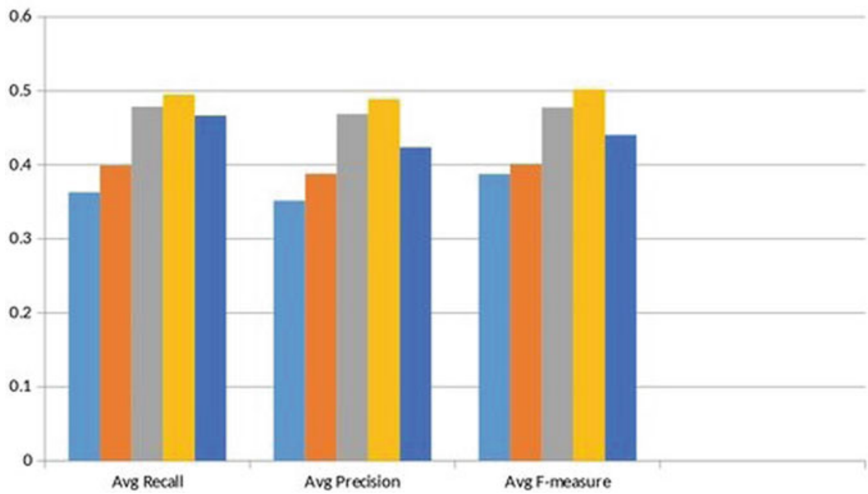


Fig. 5 ROUGE-2 scores

conducted using features in experiment 3. Turmeric colour refers to the implementation conducted using features in experiment 4, and dark blue colour refers to the proposed approach by [7].

6 Conclusion

Massive works are available in the English language, but this is not applicable for low resource languages like Assamese. It is due to the sentence structures, inflectional words, grammatical rules, etc. Also with the lack of automatic tools, unavailability of a database for ontological knowledge of words and limited scope of knowledge sharing made research in this language more challenging. Irrespective of the challenges, an approach of graph-based Assamese text summarization has been proposed in this paper based on the identification of nouns and co-sine similarity. It is seen that the use of nouns enhances the performance. The amalgamation of semantic similarity and statistical features generates a better summary in terms of sense regarding ROUGE scores for Assamese text as seen in Table 4. In the proposed system, it also outperforms the latest graph-based text summarization approach in the Assamese language.

References

1. Luhn HP (1958) The automatic creation of literature abstracts. IBM J Res Devel 2(2):159–165
2. https://en.wikipedia.org/wiki/Assamese-people. Accessed on 25 Feb 2020
3. Qasem A, Radaideh A, Bateineh D (2018) Q. A hybrid approach for Arabic text summarization using domain knowledge and genetic algorithms. Cognitive Comput 10:651–669
4. Kalita C, Saharia N, Sharma U An extractive approach of text summarization of Assamese using WordNet. In: Global WordNet conference (GWC-12)
5. Baruah N, Sarma S, Borkotokey S (2021) A single document Assamese Text summarization using a combination of statistical features and Assamese WordNet. In: Progress in advanced computing and intelligent engineering, pp 125–136
6. Baruah N, Sarma SK, Borkotokey S (2020) Evaluation of content compaction in Assamese language. Proc Comput Sci 171:2275-2284
7. Baruah N, Sarma S, Borkotokey S (2019) A novel approach of text summarization using Assamese WordNet. In: Proceedings of 4th international conference on information systems and computer networks (ISCON), pp 305–310
8. Sarma S, Medhi R, Gogoi M, Saikia U (2010) Foundation and structure of developing an Assamese WordNet. In: Proceedings of 5th international global wordnet conference (GWC 10)
9. Sobh I (2009) An optimized dual classification system for Arabic extractive generic text summarization. M.Sc Thesis: Department of Computer Engineering, Cairo University, Giza
10. Gupta V,Kaur NA (2016) Novel hybrid text summarization system for Punjabi text. Cognitive Comput 8:261–277
11. Haque M, Pervin S, Begum Z (2017) An innovative approach of Bangla text summarization by introducing pronoun replacement and improved sentence ranking. J Inf Process Syst 13(4):752–777

12. Gupta V, Lehal G (2010) A survey of text summarization extractive techniques. J Emerg Technol Web Intel 2(3):258–268
13 Prabhakaran S Cosine similarity-understanding the math and how it works (with python codes). www.machinelearningplus.com. Accessed on 2019
14. Sarma S, Bharali H, Gogoi A, Deka R, Barman AK (2012) A structured approach for building Assamese corpus: insights, applications and challenges. In: ALR@COLING, pp 21–28
15. Formulae and tables for statistical work. Edited by Rao CR, Mitra SK, Matthai Statistical Publishing Society, Calcutta (1966)
16. Lin C (2004) ROUGE: a package for automatic evaluation of summaries. In: the ACL workshop on text summarization branches out, Barcelona, pp 74–81
17. El-Haj M, Kruschwitz U, Fox C (2011) Experimenting with automatic text summarization for Arabic. Hum Lang Technol Chall Comput Sci Linguist Springer, pp 490–499
18. Hajime M, Manabu O (2000) A comparison of summarization methods based on task-based evaluation. In: Proceedings of 2nd international conference on language resources and evaluation. LREC, pp 633–639

Internet of Things (IoT) for Secure Data and M2M Communications—A Study

N. Shashikala and Monica R. Mundada

Abstract At present, the IoT has received huge amount of research considerations. The IoT is acknowledged by payless progression of information among numerous low-power embedded gadgets which utilize the Internet for interacting among others. It is anticipated that the IoT would be generally conveyed and would identify relevance in various fields of life. Requirements for IoT have recently taken in enormous attention, and connections are amped up for the trading probabilities of the information which would be created by deploying those systems. Despite what may be thought of, IoT has various protection and security worries for the last customers which limit its expansion. Because of the billions of associated gadgets, there is an incredible danger of personality and information burglary, gadget control, information manipulations, network or server alterations, and ensuing effect on application stages. Security is one of foremost innovative research-based issues that is available at present for IoT. The security contains numerous aspects—security worked inside the gadget, security of information transmissions, and information stockpiling inside the frameworks and its application. In the work done here, we have distinguished, categorized, and examined numerous security issues and best in class endeavors to determine these difficulties. The arrangement of effective security and protection conventions in IoT systems is incredibly expected to guarantee secrecy, verification, access control, and honesty, among others.

Keywords IoT security · Security threats · Privacy · M2M · IoE

N. Shashikala (✉) · M. R. Mundada
Department of CSE, Ramaiah Institute of Technology, Viswesaraiah Technological University, Bangalore, India
e-mail: shashikala.n@reva.edu.in

M. R. Mundada
e-mail: monica@msrit.edu

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_2

1 Introduction

The IoT, Internet of things, is involved billions of associated gadgets, gathering, and sharing data. This information can be assembled, analyzed, and adapted. The dynamic service environment provides novice threats concerned to security and difficulties to different ventures. The developing specialized space is emerging with the IoT. The IoT is achieving a change in perspective in administrations, foundation, and consumer businesses [1]. While this change in outlook is going on, trust and security are important prerequisites to handle various types of assaults, threats, breakdowns, and destroying effects on society. The obligation of making sure about IoT lies with gadget producing organizations and organizations that utilization the gadgets. Having a total arrangement of security terminologies is a need to sort out the risk and defeat every security issues in IoT. Few of the security necessities for IoT are presented, consisting hashing, encoding, and different types of secure conversations [2, 3]. However, more is expected to make sure about this foundation from threats and assaults just as other related interests. Propelling the innovation to make sure about the IoT condition is the inspiration of this exploration work. In order to achieve this, we (i) play out an intensive examination of the present acts of making sure about customary systems, (ii) recognize the current IoT security prerequisites and rules, (iii) investigate the IoT architecture, (iv) characterize the present dangers that IoT gadgets are confronting, and (v) address the areas of IoT. The aftereffects of this examination will be broke down and composed to give basic investigation on the present security of IoT gadgets and framework, just as to build up a total scientific categorization to give an all-encompassing perspective on safe environment of IoT. With expanded commercialization of IoT gadgets, society is turning out to be increasingly more associated with the infrastructure of the IoT—allowing society progressively powerless to the vulnerabilities of present environment of IoT. This will progressively contact our living in a larger number of ways than previous years. Therefore, researchers must handle and resolve IoT's security aspects. Traded off IoT gadgets present the danger of misusing individual data, bargaining other associated frameworks, and risks [4].

Because of quick progressions in portable communication, WSNs, and computing which are distributed in nature and RFID, exchanges with IoT gadgets have gotten more advantageous compared to as it was earlier. The IoT instruments are equipped for working together along with every other. The universe of IoT installs a tremendous assortment of instruments which incorporate developed cells, tablets, desktops, PDAs, laptops, and handier installed equipment. The IoT instruments based on sensors are practical and remotely interacting frameworks to exchange info with one another and move significant data to the systematic machine, which is centralized. The data packets from IoT instruments are additionally handled in the centralized systems and then informed to the expected goals. Along the speedy growth of interaction technologies and web innovations, the everyday tasks are increasingly concentrated on an anecdotal space of virtual universe [5].

IoT has effectively coordinated the fictional spaces and such current realities on uniform platforms. The significant concentrations of IoT are generally the setting-ups of a savvy fields and not-sure autonomous instruments, consider case, living smart, brilliant things, smartly well-being, and smarter urban zones among others [6]. In current days, the pace appropriation of the IoT instruments is exceptionally higher; an ever incrementing count of instruments are connected by means of the web. In coming days, the IoT would completely alter our styles of living and action plans. This would permit humans and instruments to interact wherever and whenever, along any equipment under ideal conditions making use of given framework and any assistance [7]. The principal goal of IoT should be to provide most superior way of living in this universe for human in coming days. Figure 1 demonstrates the selective application domain associated with IoT. Tragically, many of those instruments and software are not meant be to dealt with the protection and security assaults and that expands a great deal of privacy and security based issues in IoT instrumentations, for instance, classifications, inspections, respectability of information, access controls, secrecy, confidentiality, and so forth [8]. On continuous and consistent basis, the IoT instruments are concentrated by intruders and interlopers. An examination reveals that more than 70 percentages of the IoT instruments are exceptionally simple for hacking. Accordingly, an effectively used system is very expected to make sure about the gadgets associated with the intruders and hackers.

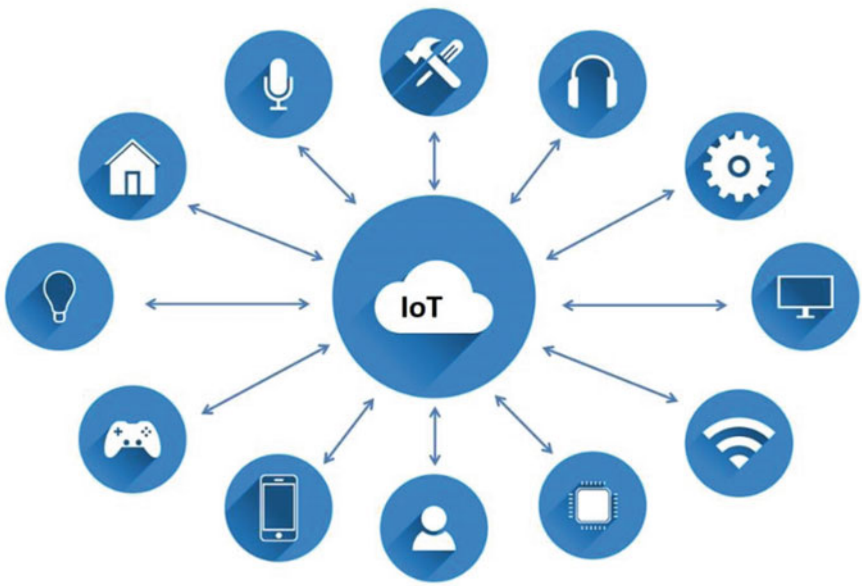


Fig. 1 Selective application domains associated with IoT

2 Related Works

In current years, with the notoriety of brilliant gadgets, the IoT has become a trendy expression to depict the smart gadget objects. This has opened up completely better approaches to advance our everyday lives in an increasingly programmed, proficient, and helpful way. As a key segment, IoT interaction is the way to IoT scenarios. In any case, because of the open idea of remote correspondence in the sensor systems and the constrained capacity including data transmission, storage, and vitality, security and protection issues are not all around tended to. It is defenseless against programmer assaults in IoT interaction. Along these lines, tending to security and protection challenges in IoT correspondence is of essential significance in IoT environment. The center explanation behind the gigantic measure of information created by Internet of things empowered gadgets is the expanding count of web empowered gadgets utilized for various purposes by people, organizations and governments. These gadgets are utilized with the end goal of information investigation, data management, creation of knowledge and its management. This helps insufficient decision making and policy management. Therefore, this huge measure of information induced by the IoT gadgets needs increasingly maximum computing capacity to process.

The work in [9] presents a structure to create and convey IoT apps in the cloud environment. The planned system profits by the present modules of Aneka cloud and moreover focuses toward novel highlights required to IoT apps. For correspondence among information origins and cloud platform Aneka, a convention MQTT which is lightweight is used. The structure proposed consists of three significant components, for example, application administrator, cloud supervisor, and information source coordination. The application supervisor is apportioned into segments like, App Monitor, App Composer, Load Balancer and Scheduler. These segments furnish the client with application side functionalities, for example, creation, planning, and observing of apps. The proposal by the researcher in [10] is increasingly engaged toward exercises of information affiliation, induction, and information disclosure process in big data management belong to IoT. The researchers additionally give précised future headings toward knowledge identification of IoT. In another work, an intellectual IoT structure has been introduced to upgrade the capacities of semantic determinations from gathered information, managing knowledge, and process for decision making and information discovery [11].

The exploration researchers in [12] introduced security-based framework for IP-empowered IoT dependent on Host_Identity_Protocol (HIP) and Datagram_Transport_Layer_Security (DTLS) adjusted to resource-based obliged gadgets. Also, researchers propose key administration framework for IoT. Protection of privacy is given by presenting Datagram_Transport_Layer_Security-pre_shared keys (DTLS-PSK) and Host_Identity_Protocol pre_shared keys (HIP-PSK) that give secure system access and correspondence. In the transaction [13], the scholars classified methodologies in handling IoT as protocol-dependent and framework-dependent methodologies, that is, presented a design-dependent security protecting system. IoT is displayed as agreeable distributed frameworks in which things coordinate to achieve person-based or helpful objectives. Contract_Net_Protocol (CNP) is stretched out to help security assurance to IoT. Most recent assessments of IoT systems which belong to security and protection could be seen in [14]. The creators concentrated on saving protection in home computerized frameworks which are professed to be stretched out to IoT apps. It is exhibited that both fundamental cryptography-related procedures and information control are utilized to spare a client against an opponent internal to the IoT system or adversary which have undermined servers available remotely.

3 Internet of Things and M2M Communications

At present, as the M2M interaction is a usually utilized term, particularly given conversation encompassing the Industrial IoT, which has a more drawn out history. The Supervisory Control and Data Acquisition (SCADA) arrangements have depended on M2M interactions for many of the decades [15], and even earlier to that the utilization of M2M correspondences permitted the utilization of ATMs, also retail location frameworks. The M2M includes direct correspondence among gadgets with no human mediation. Such correspondence could be through any channel, regardless of whether remote or wired, and the quantity of advancements, standards, and conventions for interaction is huge and developing. The correspondence might happen through a system, involving cell systems such as 3G, 4G, and GSM, or legitimately between gadgets in a point-to-point way, each consisting an alternate assault surface. A portion of the key correspondence advances incorporate RFID, Wi-Fi, Dedicated Short Range Communication (DSRC) Bluetooth, Zig Bee, and NFC. All such innovations, that change in recurrence, coverage, and range, are characterized by various norms, as analyzed in Table 1. Notwithstanding these changing interaction techniques, there available app specific norms, for example, the Meter-Bus norm, EN 13757-x, produced for the remotely perusing of power or gas meters, the ISO/IEC 14543-3 (Systems for Home Electronic), EN 50090-x (Systems for Building and Home Electronics).

We must notice the availability of an assortment of web of X depictions: the ABI explore [16] thinks about the IoT as an equal improvement to the Internet of digital and humans both, for instance. The report in [17] examines the improvement of the Internet of services, while ABB has been creating items and administrations for the Internet of things, humans, and services. The upcoming Industrial Revolutions are being created via the Industrial IoT [18], the associated vehicle motivation is forming into the Internet of Vehicles [19], and also, there are considerably progressive improvements, for example, the Internet of animal health items [20]. As of late, Qualcomm and Cisco have been pushing the utilization of the terminology IoE, Internet of everything. While few contend that the terminology might have been generated as an advertising chance by Cisco, unquestionably there are some advantages in categorizing frameworks which might go past a considerable lot of the

Table 1 Technologies for interactions used in IoT and M2M

Technology	Standard	Frequency	Coverage	Bit rate	Comments
Wi-Fi	IEEE 802.1 J	2.4/5 GHz	50 m	500 Mbps	High consumption
ZigBee	IEEE 802.15.4	2.4 GHz	100 m	250 kbps	High security
Z-Wave	ZAD12837	900 MHz ISM	50 m	40 kbps	Home automation
Sigfox	Sigfox	900 MHz ISM	10 km	1 kbps	Low consumption
Neul	Neul	458 MHz	10 km	100 kbps	Low-cost IoT
LoRaWAN	LoRaWAN	ISM bands	15 km	50 kbps	Wireless battery operated IoT
RFID	ISO/IEC J8000	LF, ISM bands	<2 m	40 kbps	
NFC	ISO/IEC J8092	1 3.56 MHz	<20 cm	424 kbps	
GSM/3G/4G	GSM, UMTS/HSPA, LTE	900/J800/ 190O/21 00 MHz	50 km	10 Mbps	High consumption
Bluetooth LE	IEEE 802.1	2.4 GHz	50 m	1 Mbps	Low consumption
6LoWPAN	RFC6282	ISM bands	n/a	n/a	
HomePlug	IEEE1901	<100 MHz	<100 m	10-500Mbps	Smart grids
Thread	Based on IEEE802.15.4	2.4 GHz	<100 m	250kbps	Up to 50 devices
DSRC	IEEE802.11p	5 GHz ISM	300 m	27 Mbps	V2V comms
WiMax	IEEE802.16	2.3, 2.5, and 3.5 GHz	10 km	10 Mbps	

ordinary employments belong to the IoT, especially taken its advancement external to M2M situations. The IoE could be considered as superset of IoT, because the M2M is viewed as subset of IoT.

The idea of the IoE unites four primary components: humans, items, process, and information, where the items are gadgets, actuators, sensors, and different things, creating information or accepting data from different origins. [21] As opposed to being confined to the human, also involve the human-made frameworks, for instance, social communities, [22] and well-being, and wellness apps. The information is broken down and prepared to make helpful data for astute decisions and for the mechanisms to control. Such idea of IoE would not simply allow assessment of the IoT as frameworks containing people and machines, yet in addition, combines the intelligence, instrumentation, and inter-networking through the processes of information [23, 24].

4 IoT Security Analysis

The quantity of IoT gadgets is expanding every single day. The purpose behind expanding the quantity of IoT gadgets is that they give comfort in human life and perform work with preferable results over people. It has been accounted for that, in 2018, the quantity of IoT gadgets dramatically multiplied since 2012, and we reached around 50 billion gadgets now, in 2020, that took a shot at the Internet. The graphical analysis in Fig. 2 demonstrates the quantity of associated IoT gadgets from 2012 till 2020. In 2025, the number of devices in internet is expected to be 1 trillion.

Table 2 demonstrates the application areas of IoT. Besides, it additionally contrasts the diverse application areas and regard to the quantity of clients, interaction technology, system size, data transfer capacity, and test beds belong to them. The Internet of things invites for thinking past customary processing. This requests small, brilliant,

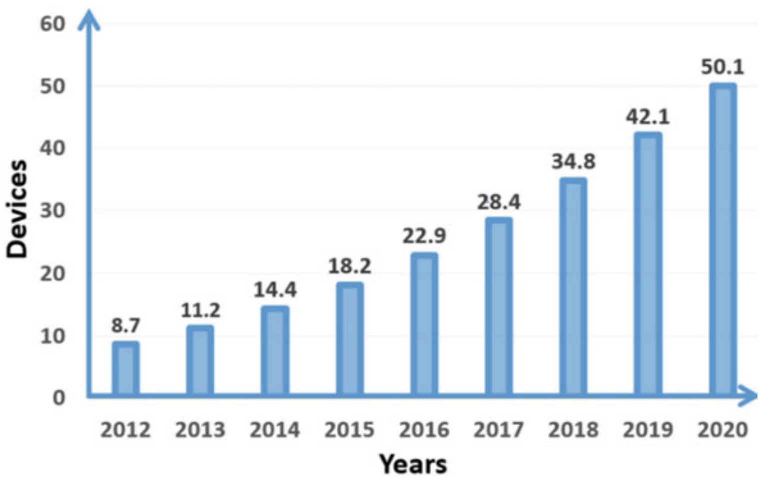


Fig. 2 Quantity of IoT devices connected to Internet from 2012 till 2020

Table 2 Comparisons of numerous application domains of IoT

	Home/Office	City	Transportation	Agriculture	Retail
Number of users	Very few	Many	Many	Few	Few
Communication	RFID and WSN	RFID and WSN	WSN	WSN	RFID and WSN
Network	Small	Medium	Large	Medium	Small
Internet	Wi-Fi, 3G, 4G	Wi-Fi,3G,4G	Wi-Fi, satellite	Wi-Fi, Satellite	Wi-Fi,3G,4G
Bandwidth	Small	Large	Medium	Medium	Small
Test beds	smart home	smart cities	Few	PSCM system	retail centers

and minimal instrument which can supplant conventional processing abilities. The wireless sensor networks, RFIDs, smart composers, cell phones, PCs, and simple to use instruments are the major advancements which will fill in as preliminary computing units for those worldwide systems. The RFIDs are also belonging to the very few of the major players in IoT strengthening innovations [25, 26], those gets into action microchips joined to any ideal object for programmatically recognitions, following and data transmissions remotely [27]. The RFIDs are made use in benefits of the supply chain management, monitoring of ports and retails.

Interestingly with all illustrated challenges, the graphical analysis in Fig. 3 demonstrates the diverse IoT information managing challenges, that are spoken to on a level plane, while non-horizontal lines describe the quantity of information managing models appeared in the literatures. When experiencing the literatures, it was seen that there was lesser researches on the vast majority of the information managing issues, for example, data analysis, and information storage and data aggregation. Then again, there was even lesser researches on zones, for example, information security, creation of knowledge, context management, and information heterogeneity. Figure 3 presents this connection of information management models and their works toward various information management issues.

The Sonic Wall investigations report about IoT malware assaults bounced 215.7% till 32.8 million out of 2018 (10.4 million out of 2017 end). The initial two-fourth of 2019 have just outpaced the initial two-fourth of 2018 by much above 55%. In the event that this rate proceeds, that would be one more record-setting year for malware assaults of IoT. The graphical analysis in Fig. 4 illustrates the Sonic Wall data about increase of IoT malware attacks in 2018, and this action progressed in first quarter of year 2019.

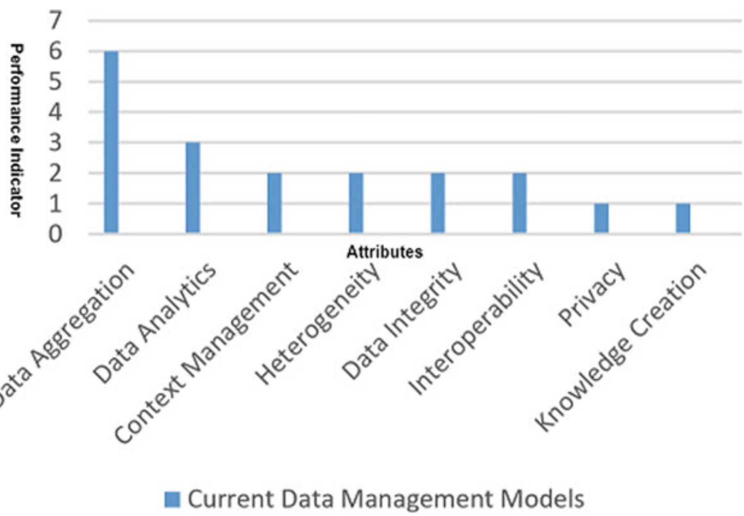


Fig. 3 Present contribution toward recognized IoT

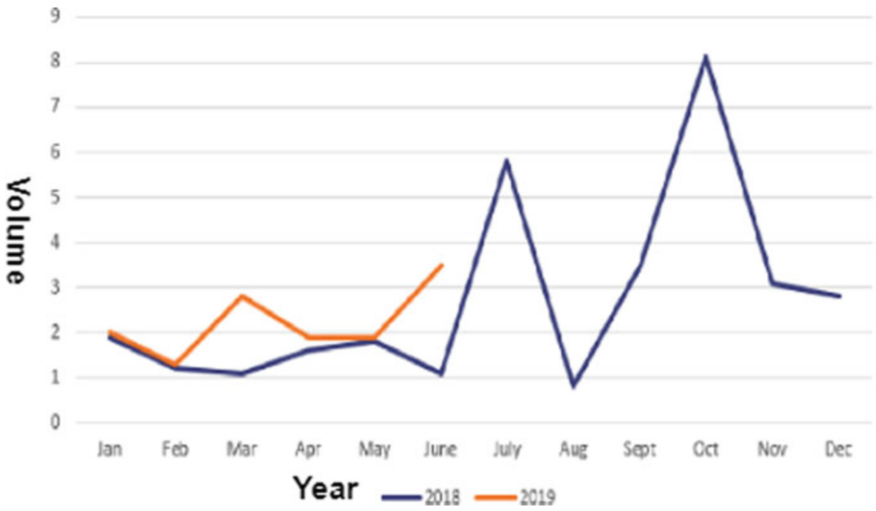


Fig. 4 IoT malware attacks increased in 2018 and 2019

In past times worth remembering, the clients were high up in the natural way of life. The liabilities were overseen and restricted by the contacts. The dependable liability measurements are applied to survey dangers for devices/software failure. The end users, buyers not figured in to risk assessments. In today's advancements, the data accessibility from storage industry products now directly affects all end users (customers, consumers, individuals, industry, and municipalities). The liability is open ended along with many flavors. No reliable liability metrics could be applied to assess risks for device/software failure. Consider the data accessibility failures such as management console breach, access to data stored is blocked, device and sensors cannot call home or each other, IoT ecosystem balance malfunctions or fails. Then the affected communities will be end users, hospitals, medical facilities, big data analytics providers who monetize IoT data.

5 Analysis of Various Categories of Attacks and Suggested Solutions

The Internet of things is confronting different kinds of assaults consisting passive and dynamic attacks which may handily distract the usefulness and cancel the advantages of its functionalities. As per the passive assault, an intruder simply senses the hub or might take the data; however, it never assaults genuinely. Nonetheless, the dynamic assaults distract the performances genuinely. Those dynamic assaults are organized into two more classes which are the attacks happens internally and externally. Such powerless attacks could secure the gadgets for smart communications. Subsequently, the safety limitations should be installed to keep instruments from malignant attacks.

The behavior and levels of risks associated in various kinds of attacks are discussed. Various degrees of attacks are ordered into four kinds as per their conduct and also present potential solutions for the attacks.

i. Attack level-one: When an intruder attempts to hack the systems and these hacking is not successful.
ii. Attack level-two: When a hacker or a spy is generally tuning in to the system medium however do not change the information integrity.
iii. Attack level-three: When a hacking is proceeded on the systems and it changes the integrity of information or alters the contents.
iv. Attack level-four: When a hacker attacks on systems by taking unauthorized access and making out an illicit activity, allowing the systems to be inaccessible, transferring bulk messages, or performing system jam.

The information given in Table 3 shows various kinds of attacks, levels of threats, behavior of attacks, and also suggested solutions to track those attacks. The attack level-one is considered as low-level assault, attack level-two as medium-level assaults, attack level-three as high-level assaults, and attack level-four is considered as extremely high-level assaults.

The concerns for the security in the IoT are the biggest issues. The app information of IoT can be enterprise related, industry related, and consumer related. Such app information must be made secured and should stay confidential in terms of data stealing and altering. Considering the instance, the IoT apps might save the outcomes of a person's well-being or procuring mart. The IoT boosts the interactions between instruments; still, there are challenges detected with the versatility, reaction time, and accessibility. The privacy and secure transmission of data in the Internet should be major concern.

Among various challenges related to security, the much significant challenges applicable to IoT are taken for discussion.

i. Privacy of information: Some of the vendors gather the information of their customers for the analysis. This challenges privacy of information while transmission.
ii. Security of information: When the data transmission is done continuously, we must be able to hide the data from spywares in the Internet.
iii. Common standard lacking: As there will be numerous standards related to IoT gadgets and manufacturers, this is a bigger inspection to detect allowed and non-allowed instruments connected with the web.
iv. Concerns on technical aspects: Because of the expanded usage of IoT instruments, the traffic created by those instruments is likewise getting expanded. Consequently, there is a requirement to expand limit of the system; in this manner, it is additionally an inspection to store the tremendous measure of information for examination and further final stockpiling.

Table 3 Various kinds of attacks, levels of threats, behavior of attacks, and suggested solutions to track those attacks

Type	Threat level	Behavior	Suggested solution
Passive	Low	Usually breach data confidentiality. Examples are passive eavesdropping and traffic analysis. Hostile silently listen the communication for his own benefits without altering the data	Ensure confidentiality of data and do not allow, an attacker to fetch information using symmetric encryption techniques
Man in the middle	Low to medium	Alteration and eavesdropping are the examples of this attack. An eavesdropper can silently sense the transmission medium and can modify the data if encryption is not applied and steal the information that is being transmitted. Hostile may also manipulate the data	Apply data confidentiality and proper integration on data to ensure integrity, Encryption can be also applied so that no one can steal the information or modify the information or encode the information before transmission
Eavesdropping	Low to medium	The information content may be lost by an eavesdropper that silently senses the medium. For example in medical environment, privacy of a patient may be leaked	Apply encryption on all the devices that perform communication
Gathering	Medium to high	Occurs when data is gathered from different wireless or wired medium. Examples are skimming, tampering, and eavesdropping. Data is being collected to detect messages. Messages may also be altered	Encryption can be applied to prevent this kind of attack. Identity based method and message authentication code can also be applied in order to prevent the network from such malicious attacks
Active	High	Effects confidentiality and integrity of data. Hostile can alter the integrity of messages, block messages, or may re-route the messages. It could be an internal attacker	Ensure both confidentiality and integrity of data. To maintain data confidentiality, symmetric encryption can be applied. An authentication mechanism may be applied to allow data access to only authorized person

(continued)

Table 3 (continued)

Type	Threat level	Behavior	Suggested solution
Imitation	High	It impersonate for an unauthorized access. Spoofing and cloning are the examples of this attack. In spoofing attack a malicious node impersonate any other device and launch attacks to steal data or to spread malware. Cloning can re-write or duplicate data	To avoid from spoofing and cloning attacks, apply identity based authentication protocols. Physically unclonable function is a counter measure for cloning attack
Privacy	High	Sensitive information of an individual or group may be disclosed. Such attacks may be correlated to gathering attack or may cause an imitation attack that can further lead to exposure of privacy	Apply anonymous data transmission. Transmit sample data instead of actual data. Can also apply techniques like ring signature and blind signature
Interruption	High	Affects availability of data. This makes the network unavailable	Applying authorization, only authorized users are allowed to access specific information to perform certain operation
Routing diversion	High	Only the route is diverted showing the huge traffic and the response time increased	Ensure connectivity based approach so no route will be diverted
Blocking	Extremely high	It is type of DoS, jamming, or malware attacks. It sends huge streams of data which may leads to jamming of network, similarly different types of viruses like Trojan horses, worms and other programs can disturb the network	Turn on the firewall, apply packet filtering, anti-jamming, active jamming and updated antivirus programs in order to protect the network from such attacks
Fabrication	Extremely high	Affects the authenticity of information. Hostile can inject false data and can destroy the authenticity of information	Data authenticity can be applied to ensure that no information is changed during the transmission of data

(continued)

Table 3 (continued)

Type	Threat level	Behavior	Suggested solution
DoS	Extremely high	Malicious user may modify the packets or resend a packet attain and again on network. User can also send bulk messages to devices in order to disturb the normal functionalities of devices	Apply cryptographic techniques to ensure security of network. Apply authenticity to detect the malicious user and block them permanently. In this way, the network is prevented from damage

Table 4 Category-wise installation of IoT units in millions

Category	2013	2014	2015	2020
Automotive	96	189.6	372.3	3511.1
Consumer	1842.1	2244.5	2874.9	13,172.5
Generic Business	395.2	479.4	623.9	5158.6
Vertical Business	698.7	836.5	1009.4	3164.4
Grand total	3032	3750	4880.5	25,006.6

Up till now, there has been tons of tasks performed in the environment belong to the IoT security. The concerned tasks could be categorized into security related to apps, security related to networks, security related to systems, and so on. Table 4 demonstrates the units of IoTs installed in various categories of businesses. It shows the drastic increase in the devices connected to Internet of things from 2014 till 2020.

The observation made in the graph showed in Fig. 6 gives us clear picture of quantity of devices connected to Internet of things that drastically change every year. The grand total gives in 2020 is more than all together connected units in previous years. The security for execution of the applications, maintaining the system and network capability in the Internet will be handled according to the suggested solutions provided above.

6 Conclusions

During the days ahead, the IoT will act an indispensable role and will alter the living styles, working models, and standards. The utilization of IoT in various apps is required to grow rapidly in the years ahead. The IoT allows billions of instruments, people groups, and functionalities to associate along others for trading data. The preliminary accentuation of the work done was to characterize most important issues related to security of IoT majorly, cantering the security assaults and their counter effects. Due to the lack of instrument related to security in IoT devices, various IoT devices become simple objectives, and also, this is not in victim's information on

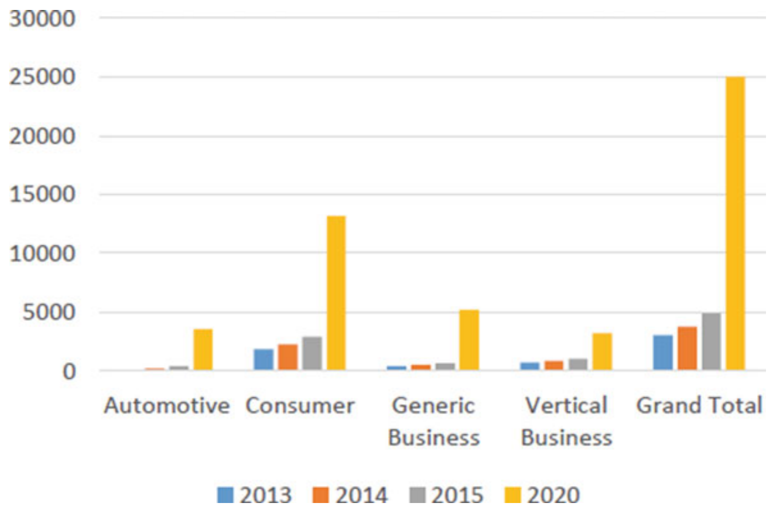


Fig. 6 Graphical demonstration of increase in IoT connected units

being tainted. The security necessities are considered for discussions, for instance, trustworthiness, confidentiality and authorization, and so on. The attacks are classified as level-one (low-level), level-two (medium-level), level-three (high-level), and level-four (extremely high level) attacks alongside their behavior and also recommended solutions for practicing these attack. Consideration the importance of IoT apps-related security, it is highly critical to present techniques for security in IoT instruments and interaction systems. Additionally, to shield from any intruders or threats concerned to security, also prescribed not to make use of passwords which are default for the instruments and follow the security essentials for the devices well before the usage in the beginning. Later on, we might want to make an anxious and complete rundown of IoT gadgets and make a dashboard or portable app to accomplish more prominent consciousness of controls of security which should be executed on different IoT instruments for a progressively secure experience of the customer.

References

1. Z Yan P Zhang A Vasilakos 2014 A survey on trust management for Internet of Things J Netw Comput Appl 42 120 134
2. Granjal J, Monteiro E, Sa Silva J (2015) Security for the internet of things: a survey of existing protocols and open research issues. IEEE Commun Surv Tutor 17:1294–1312
3. Ambrosin M et al (2016) On the feasibility of attribute-based encryption on internet of things devices. IEEE Micro 36(6):25–35. https://doi.org/10.1109/MM.2016.101
4. Kharchenko V, Kolisnyk M, Piskachova I, Bardis N (2016) Reliability and security issues for IoT-based smart business center: architecture and Markov model. In: 2016 third international conference on mathematics and computers in sciences and in industry (MCSI), Chania, pp 313–318

5. Kumar JS, Patel DR (2014) A survey on internet of things: security and privacy issues. Int J Comput Appl 90(11)
6. Abomhara M, Køien GM (2014) Security and privacy in the internet of things: current status and open issues. In: International conference on privacy and security in mobile systems (PRISMS). IEEE, pp 1–8
7. L Atzori A Iera G Morabito 2010 The internet of things: a survey Comput Netw 54 15 2787 2805
8. Hossain MM, Fotouhi M, Hasan R (2015) Towards an analysis of security issues, challenges, and open problems in the internet of things. In: 2015 IEEE world congress on services (SERVICES). IEEE, pp 21–28
9. Khodadadi F, Calheiros RN, Buyya R (2015) A data-centric framework for development and deployment of internet of things applications in clouds. In: 2015 IEEE tenth international conference on intelligent sensors, sensor networks and information processing (ISSNIP). IEEE
10. R Roman J Zhou J Lopez 2013 On the features and challenges of security and privacy in distributed internet of things Comput Netw 57 10 2266 2279
11. Sicari S et al (2015) Security, privacy and trust in internet of things: the road ahead. Comput Netw 76:146–164
12. Garcia-Morchon O, Keoh SL, Kumar S, Moreno-Sanchez P, Vidal-Meca F, Ziegeldorf JH (2013) Securing the IP based internet of things with HIP and DTLS. In: Proceedings of sixth ACM conference on security and privacy in wireless and mobile networks, Budapest, Hungary
13. A Samani HH Ghenniwa A Wahaishi 2015 Privacy in Internet of things: a model and protection framework Proc Comput Sci 52 606 613
14. Schurgot MR, Shinberg DA, Greenwald LG (2015) Experiments with security and privacy in IoT networks. In: Proceedings of IEEE world of wireless, mobile and multimedia networks (WoWMoM), Boston, MA
15. J Morrish 2014 Business models for machine-to-machine (M2M) communications C Anton-Haro M Dohler Eds Machine-to-machine (M2M) communications: architecture, performance and applications Wood head Publishing Oxford 339 353
16. Maple C (2017) Security and privacy in the internet of things. ABI Research 155–184, Published online: 24 Aug 2017
17. P Buxmann T Hess R Ruggaber 2011 Internet of services Bus Inf Syst Eng 1 5 341 342
18. Sadeghi A-R, Wachsmann C, Waidner M (2015) Security and privacy challenges in industrial internet of things. In: 52nd ACM/EDAC/IEEE design automation conference, San Francisco, CA, June 8–10. ACM Press, New York, pp 1–6
19. Gerla M, Lee E-K, Pau G, Lee U (2014) Internet of vehicles: from intelligent grid to autonomous cars and vehicular clouds. In: 2014 IEEE world forum on internet of things (WF-IoT), 12(9):241–246. Seoul, Korea, March 6–8
20. Smith D, Lyle S, Berry A, Manning N, Zaki M, Neely A (2015) Internet of animal health things (IoAHT) opportunities and challenges. University of Cambridge
21. Ahmed SST, Thanuja K, Guptha NS, Narasimha S (2016) Telemedicine approach for remote patient monitoring system using smart phones with an economical hardware kit. In: 2016 international conference on computing technologies and intelligent data engineering (ICCTIDE'16). IEEE, pp 1–4
22. Ahmed SST, Sandhya M, Shankar S, ICT's role in building and understanding Indian telemedicine environment: a study. In: Information and communication technology for competitive strategies. Springer, pp 391–397
23. Bojanova I, Hurlburt G, Voas J (2014) Imagineering an internet of anything. Computer 47(6):72–77. https://doi.org/10.1109/MC.2014.150
24. Kumar SS, Ahmed ST, Vigneshwaran P, Sandeep H, Manjunath Singh H (2021) Two phase cluster validation approach towards measuring cluster quality in unstructured and structured numerical datasets. J Ambient Intel Human Comput 12(7):7581–7594
25. Gubbi J et al (2013) Internet of things (IoT): a vision, architectural elements, and future directions. Future Gen Comput Syst 29(7):1645–1660

26. Al-Fuqaha A et al (2015) Internet of things: a survey on enabling technologies, protocols, and applications. IEEE Commun Surv Tutor 17(4):2347–2376
27. Chen C-M et al (2012) RCDA: recoverable concealed data aggregation for data integrity in wireless sensor networks. IEEE Trans Parallel Distrib Syst 23(4):727–734
28. S Chen H Xu D Liu B Hu H Wang 2014 A vision of iot: applications, challenges, and opportunities with china perspective IEEE Internet Things J 1 4 349 359

Development of Walking Assistants for Visually Challenged Person

Aditya Sandeep Lokare, Prasanna Venkatesh, S. Vela Sitthanathan, and T. Mohanraj

Abstract With the development of technology, many inventions have been made which have helped to make the lives of differently abled people easier. Prosthetic arms and legs have been developed, hearing aids are now readily available, glasses and contact lenses are available for people who have myopia or hypermetropia, motor-operated wheelchairs are available for people with impaired legs. Most of the benefits of technology advancement have little consideration for the visually impaired even though they constitute about 3.6% of world's population. However, with the advent of artificial intelligence, machine learning and the Internet of things, different types of helping aids have been developed to facilitate a visually challenged person to navigate. Unfortunately, these helping aids either have limited scopes and too many constraints or are very expensive. The device intends to assist a visually impaired person to walk around by integrating machine learning algorithms and image processing techniques.

Keywords Machine learning · Image processing · IoT · AI

1 Introduction

The visually impaired people face challenges in workstations [1], crossroads, stairs and their day-to-day life [2]. The development in technology has resulted a range of blind man assistive devices which facilitate the navigation of the visually challenged [3–5]. With the advancement in technologies, ideas considered unfeasible can be re-attempted and existing devices can be further improved [6, 7]. An idea was developed to produce a device that could be affordable by every blind person, which will effectively help them navigate their way through their day-to-day life. The wearable device comes with a Jacket fitted with processors, a wearable helmet with a camera and a pair of headphones. The device offers to help the user by providing constant

A. S. Lokare · P. Venkatesh · S. V. Sitthanathan · T. Mohanraj (✉)
Department of Mechanical Engineering, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, Coimbatore, India
e-mail: t_mohanraj@cb.amrita.edu

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_3

feedback through the earphones, which are processed by the processors, with the help of the camera that provides live video feed to the processor.

The data is then fed to the image processing software where the objects are identified and the distance of the object from the person is estimated. The device then issues warnings, voice output, based on the data given by the machine learning algorithm, and also suggests the route the person must take in the form of voice messages. This helps in making sure that the blind person knows what is in his immediate surrounding, as well as being provided with a safe path with the help of the audio prompts.

2 Literature Review

Machine learning has a variety of applications such as face recognition, predictive maintenance [8–10], automated sorting and condition monitoring [11–13]. Image processing and machine learning working in tandem are, however, a relatively new concept [14–16]. Face recognition is using a camera input, which is later grayscale and then fed into Gabor algorithm was found to be effective [17]. The paper used a database with multiple pictures of faces stored in the form of a.py file to confirm or identify the faces. The histogram approach which was used as an efficient way of detecting faces.

The attempt to help visually impaired people read, Megha Arakeri [18], describes using a camera with Raspberry Pi processor using Google 3D space analyser to identify the alphabets and effectively read the content held in front of the camera. The paper emphasizes that the output provided an audio/voice output. Sandeep Kumar [19] describes a portable device made using a Raspberry Pi processor which can overcome the glitches of identification of text in varying brightness of illumination and also describes a way to help the camera take effective pictures in low illumination.

As said by Mann [20], object detection can be classified as sensor-based, computer vision-based and smartphone-based devices. An example of sensor-based device is given by Elmannai [21], while Hakobyan [22] discusses mobile-based technology in detail. The survey conducted, by Min [23], on 30-old-aged visually impaired users found that the most desired qualities of the assisting device are mobility and balance of the user. The output recommended is an audio output, which is further reinforced by the claims made by Massof [24] and Scheggi [25]. Palanisamy [26] implemented distance measurement using ultrasonic sensor and a micro-controller. The paper describes how placing the sensor at an angle of 45° helps in the detection of obstacles. Being fitted with an IR sensor the device is apt for detecting fire hazards too, with only the range of detection being a limiting factor. Dhanuja [27] talks about a sensor-based device that uses Arduino (ATmega328 microcontroller) for microcontroller, the mode of inputs being ultrasonic and IR sensors and use of a voice playback module as output. The device classifies distances as ‘closer to the user’, ‘near the user’ and ‘far away’. Based on the classifications, the user is prompted a path to navigate around the obstacles. Ultrasonic sensor was also used by Aymaz [28] and was proved to have a wide range of obstacle detection.

Object detection using image processing and machine learning was implemented by Herghelegiu [29]. The device identifies cliffs, pits ditches and other surface irregularities using a camera and prompts the user by means of an audio output or a tactile braille script and the effectiveness of vibration and tactile output. Kamal [30] provides an in-depth study on advantages and disadvantages of currently available devices. It also shows that the blind man assisting device can be made using a camera and sonar sensors on either side of the user working together to help judge distances and a vibration output to prompt the user about his/her course of action.

Velázquez [31] conducted a research on the optimum position for the placement of the camera for the assistance devices. The paper concludes that the helmet is the most optimum position for the camera. Min [17] analyses the merits and demerits of the existing computer vision-based devices and provides an insight on how future implementation of IoT technology will further make the assistance devices more interactive and effectively prompt the user to avoid obstacles. Priya [32] uses Open CV to detect objects and pre-recorded voice output to help the user navigate. The algorithm used is YOLO, which is state-of-the-art object detection algorithm, thus reducing the processing time to 1/20 of a second.

3 Methodology

3.1 Material Selection

To make an assisting device, an in-depth user-centric research is necessary so as improve the user's experience, the research can be in form of surveys, practical tests, etc. [33–35]. The needs of the visually impaired people were understood from the one–one interactions, and based on the needs, the device was divided into independently functioning units or modules.

The modules are:

(1) Sensing module
(2) Processor
(3) Output module.

For the device to be efficient, it is imperative that the components used in its making be optimal and can satisfy the requirements of each of the module. Based on the requirements and the literature reviewed, the potential components which could cater to the needs were selected. These potential components were rated based on its ability to satisfy the needs of the system and then re-rated after assigning priority to each of the quality. The pros and cons of each component were rated as $+1$ and -1, respectively; neutral was represented as 0. The component with the highest grades was finally selected as the component to be used in the system. The same process was repeated to find the optimal locations or positions for the respective components. **Sensing Module**. The sensing module plays a very important role in object detection as, the input for the process is provided by the sensor; i.e. the quality of the input will directly be responsible for the quality of the output. For object detection to work properly, it is imperative that the image input is of high quality and is easily identifiable; otherwise, it would lead to time delays or even erroneous judgement which in some cases could be fatal. The sensing module is responsible for sensing objects in the direct vicinity of the user, convert the data in a form which can work with the processor.

The basic requirements of the sensing module are

(1) High picture quality (at least 1080p).
(2) Economically feasible (price < 2500).
(3) High Sn ratio.
(4) High area of coverage (should cover 360*).

The components shortlisted after taking the requirements into consideration and reviewing the literature related to the module were

(1) Raspberry piV1 (5 mp camera).
(2) Raspberry pi V2 (8 mp camera).
(3) Raspberry pi HQ (12 mp camera).

After the rating and ranking, it was found that Raspberry Pi model V1 was the most suitable. The selection of the component was followed by the positioning of the component, and the optimal position was found out to be the wearable helmet.

Processor. The data input from the sensing device needs to be converted into useful information so that the images could be interpreted and objects in the surroundings could be identified. Most of the work in the device is done by the processor, which includes image processing, distance calculation, identifying the directions, object detection, finding ways to navigate around the obstacles, and providing the appropriate output. To perform all these tasks, a robust processor is required. The processor should be light (in terms of weight) as it would be carried by the visually impaired person. The processor should also be compatible with machine learning algorithms and should be able to run a variety of programming languages. Based on the literature reviewed, the processor to be used was found out to be Raspberry Pi 3B owing to its robust design, ability to execute complex programs and compatibility with machine learning.

A wearable backpack was found to be the most effective position for the processer.

Output Module. The processor interprets the data, and the obstacles around the user are detected, but the user would require the device to prompt him/her for further actions, which needs to be soothing and the user can comprehend. This set of instructions should be clear and would help the user understand the surroundings based on the output provided. Prospective components chosen after shortlisting from the list of all possible output devices were

(1) Vibration motor
(2) Heat generator
(3) Recorded voice.

When selecting ideal components to be used in the making of the device, it was equally important that these components are positioned so as to enhance the effectiveness of the components and making up for its defects. So, the rating and ranking were done to select the positions for the components.

Position of the sensing module. The position of the sensing module decides the range and the effectiveness of the device. If the position of the sensing element is not apt, it might lead to reduction in range, unstable images or obstructed view, it may also cause damage to the module. These flaws can be very fatal when the user is outdoors; therefore, a set of basic requirements were agreed upon, and the potential positions were rated based on those requirements.

The requirements are:

(1) Images/videos should be stable.
(2) Objects should not obstruct the field of view.
(3) The camera should not be in a position susceptible to damage.
(4) Additional cost incurred due to the positioning of the camera should be minimum.
(5) The camera should not restrict motion of the user or cause any additional inconvenience.

Position of the processor. The processor being the brain of the device is required to be placed in a safe location not too far off from the sensors, as long-distance transmission might cause noise or delays. The factors considered during the selection of the position are

(1) Effectiveness.
(2) Additional costs incurred to be minimum.
(3) The position of the processor should not restrict the motion of the user.
(4) The positioning of the processor should not cause inconvenience to the user.

Position of the output device. The output device is the only way of communication between the device and the user; it is the means by which the device warns the user of potential dangers and helps the user to picturize the world around himself. So, it is imperative that the output device is positioned in a way to help the user understand the signals easily.

Criteria chosen for the output device are:

(1) Effectiveness—Is the position going to enhance the output of the device.
(2) Additional costs incurred to be minimum.
(3) The output device should not restrict the motion of the user.
(4) The positioning of the output device should not cause inconvenience to the user.
(5) The risk of damage to the output device must be minimum at that position.

Table 1 Selection of output model and camera

Parameters	Priority		Camera		Output module		
		Module V1	Module V2	Module HQ	Motor	Heat	Voice
Cost	4.5	0	−4.5	−4.5	0	0	0
Size	4	0	−4	−4	0	0	0
Weight	4	0	0	0	0	0	0
Resolution	5	0	0	5	0	0	0
S/N ratio	2	0	0	1	0	0	0
Fixed focus	1	0	0	0	0	0	0
Coverage angle	3	0	1	2	0	0	0
Maintenance	2	0	0	0	0	−2	2
Response time	5	0	0	0	0	−5	−5
Battery life	3	0	0	0	0	−4	4
Health effects	3	0	0	0	0	−3	0
Total		0	−1	−1.5	0	−14	1

Table 2 Deciding the position of the camera, processor and output module

Parameters	Priority	Camera			Output module			Processor		
		Helmet	Jacket	Cane	Helmet	Jacket	Neck	Bag	Belt	Vest
Cost	4	0	−1	0	0	−1	1	0	0	−4
Effectiveness	5	0	−1	−1	0	−5	2	0	0	0
Movement	4	0	1	−1	0	−2	0	0	−3	0
Protectivity	2	0	1	1	0	4	−1	0	0	0
Restrictions	0	0	0	0	0	−1	1	0	−2	−2
Total		0	0	-1	0	−4	−1	0	−5	−6

Table 1 shows the selection process of the camera and the output module as they were rated and ranked among the possible options. Table 2 shows the decision-making process of placement of the modules on the user.

3.2 *Programming and Calculation*

This section would focus on the software part of the project ranging from the object detection algorithm to the distance measurement.

Algorithm. After the selection of hardware, the software was decided upon. The device needed to be fast, and it was necessary that it detects the objects accurately. In terms of algorithm, there were no other competitions to the state-of-the-art object detection algorithm YOLO and is proved to be helpful for pedestrian environment

[36]. The accuracy was great, and the gap in speed couldn't be any more pronounced at 30FPS [37] (the counterparts had a FPS of around 12). Other object detection algorithms like CNN, RCNN, Fast RCNN are very heavy on the processor and are slow when compared to YOLO. The algorithm can also run-on systems with no GPU [38], thereby reducing the restrictions on the processor. The algorithm can be further trained to detect a wider range of objects, thereby increasing the scope of detection [39]. The algorithm divides the whole picture into a grid and then assigns the probability of an object being present in each box, if the box doesn't have an object, it ignores the box, thus saving a lot of computational time. Figure 1 shows the object detection along with the probabilities.

Distance Measurement. Based on the camera specifications, the number of pixels the image takes and the size of the object and the distance from the object can be calculated by the formula in Eq. 1.

$$\text{Distance} = \begin{pmatrix} \text{Focal length[mm]} \times \text{Real height[mm]} \\ \times \text{Image height}\big[\text{pixels}\big] \end{pmatrix} \div \left(\text{Object height[mm]} \times \text{Sensor height}\big[\text{pixels}\big]\right) \tag{1}$$

After the distance is calculated, the device prioritizes the closest objects and then gives them directions to move around the obstacles.

Fig. 1 Object detection in YOLO

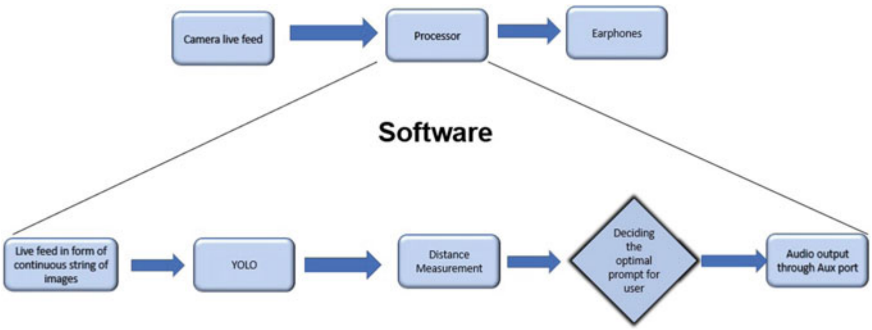


Fig. 2 Working of the device

3.3 Working of the Device

The camera captures the video live feed and sends it to the Raspberry Pi processor, where it is taken image by image and fed to the YOLO algorithm. The program then estimates the distance of the objects from the device and then decides on an appropriate path or prompt for the user. The prompt is then relayed to the user via the earphones around the neck of the user. Figure 2 shows the processes involved in working of the device.

4 Results and Discussion

The device, set-up shown in Fig 3, is designed to monitor the surroundings and vocally explain the surroundings to a visually impaired person. The device successfully helps the person navigate by prompting him using recorded voice messages.

Though not fully perfected, with the height restriction of 2 feet above ground, we intend to resolve the restriction so that the blind person can be prompted about the unevenness of the ground. With a few modifications in the output mode, i.e. using vibration signals instead of voice messages, we intend to make this device available for the visually impaired who are deaf. Figure 4 shows the device identifying the objects and calculating the distance from the objects. Table 3 shows the sample of objects detected and distances measured which were later verified with the actual distances of the objects from the camera.

The device is a practical tool to describe the surroundings to the visually impaired; at the moment, it is recommended that the device to be used as an additional support to the pre-existing cane to help the visually impaired. The device can give out directions but the unevenness of the ground and variations in the surface of the ground like speed breakers and stones need the cane to be detected. The YOLO algorithm used has a high FPS (30 FPS) enabling the device to detect sudden changes and keep up with the walking user. In terms of speed, YOLO has no competition terms of detection

Fig. 3 Device set-up

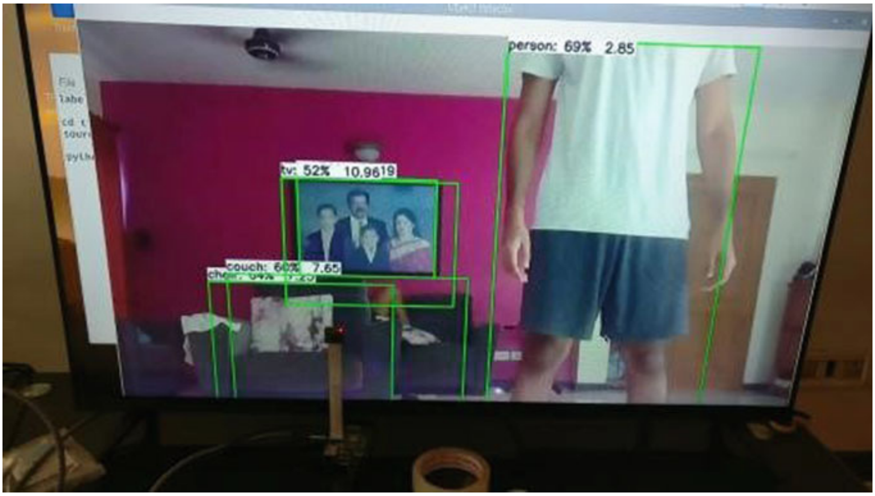


Fig. 4 Detection of common objects and distance measurement

efficiency; however, YOLO is incapable of detecting very small objects like nails and small coins, which can be detected by Faster RCNN, but it must be noted that the advantage gained in speed is more than enough to make the users turn a blind

Table 3 Object detection and distance measurement sample outputs

Original object	Object detected	Distance estimated	Original distance
Cell phone	Cell phone	3.1 ft	3 ft 1”
Person	Person	4.6 ft	4 ft7”
Keyboard	Keyboard	1.4 ft	1 ft 5”
Apple	Apple	0.6 ft	19 cm
Chair	Chair	4.1 ft	4 ft
Suitcase	Suitcase	3.3 ft	3 ft 4”
Cupboard	Refrigerator	6 ft	5 ft 11”
Refrigerator	Refrigerator	5.04 ft	5 ft
Bottle	Bottle	1.6 ft	1 ft 7”
Clock	Clock	5.01 ft	5 ft

eye towards the small objects missed by the device [39, 40]. The device is made so that the person can navigate through a moderately crowded street.

5 Conclusion

There are many assistive devices in the open market, but most of them have limited scope and are too constrained to be used. This device is a cost-effective way to overcome the restrictions and can be used in practical situations. Although the device is not completely flawless, it is the most practical solution available for almost one-tenth the price of the assistive technologies available in the market. The device is priced at one-fourteenth the price of its counterparts at Rs. 8750 and is practical enough for the user to move around in simple environments confidently. The device uses YOLO algorithm with an FPS of 30 FPS which is phenomenal when compared to other algorithms, thus enabling the device to be fast and effective. In future, the device can be made more compact and the overall parameters of the device could be upgraded making it an ideal tool used by the visually impaired and possibly replacing the walking cane which has been used for decades.

References

1. Wolffe KE, Roessler RT, Schriner KF (1992) Employment concerns of people with blindness or visual impairments. J Vis Impair Blind 86:185–187
2. Murphy E, Kuber R, McAllister G, Strain P, Yu W (2008) An empirical investigation into the difficulties experienced by visually impaired Internet users. J Univ Access Inf Soc 7:79–91
3. Xiao J, Joseph SL, Zhang X, Li B, Li X, Zhang J (2015) An assistive navigation framework for the visually impaired. IEEE Trans Human-Mach Syst 635–640
4. Hersh M, Johnson MA (2008) Assistive technology for visually impaired and blind people

5. Islam MM, Sadi MS, Zamli KZ, Ahmed MM (2019) Developing walking assistants for visually impaired people: a review. IEEE Sens J 19(8):2814–2828
6. Hu M, Chen Y, Zhai G, Gao Z, Fan L (2019) An overview of assistive devices for blind and visually impaired people. J Robo Auto 34:580–598
7. Tapu R, Mocanu B, Zaharia T (2020) Wearable assistive devices for visually impaired: a state of the art survey. Pattern Recogn Lett 137:37–52
8. Yuvaraju EC, Rudresh LR, Saimurugan M (2020) Vibration signals based fault severity estimation of a shaft using machine learning techniques. Mat Today Proc 24:241–250
9. Saimurugan M, Praveenkumar T, Krishnakumar P, Ramachandran KI (2015) A study on the classification ability of decision tree and support vector machine in gearbox fault detection. J.A.M.M vol 813. Trans Tech Publications Ltd., pp 1058–1062
10. Saimurugan M, Ramachandran KI (2014) A comparative study of sound and vibration signals in detection of rotating machine faults using support vector machine and independent component analysis. Int J Data Anal Tech Strat 6(2):188–204
11. Mohanraj T, Yerchuru J, Krishnan H, Aravind RN, Yameni R (2021) Development of tool condition monitoring system in end milling process using wavelet features and Hoelder's exponent with machine learning algorithms. Measurement 173:108671
12. Sreenath PG, Praveen Kumare G, Pravin S, Vikram KN, Saimurugan M (2015) Automobile gearbox fault diagnosis using Naive Bayes and decision tree algorithm. In: Applied mechanics and materials, vol 813. Trans Tech Publications Ltd., pp 943–948
13. Praveenkumar T, Saimurugan M, Krishnakumar P, Ramachandran KI (2014) Fault diagnosis of automobile gearbox based on machine learning techniques. Proc Eng 97:2092–2098
14. Terven JR, Salas J, Raducanu B (2013) New opportunities for computer vision-based assistive technology systems for the visually impaired. J Comp 47:52–58
15. Bhowmick A, Hazarika SM (2017) An insight into assistive technology for the visually impaired and blind people: state-of-the-art and future trends. J Multimodel User Interfaces 11:149–172
16. Nalini A (2018) Smart walking assistance device with embedded control system. IJPAM. 120(6):137–145
17. Al-Muqbali F, Al-Tourshi N, Al-Kiyumi K, Hajmohideen F (2020) Smart technologies for visually impaired: assisting and conquering infirmity of blind people using AI technologies. In: 2020 12th annual undergraduate research conference on applied computing (URC). IEEE, pp 1–4
18. Arakeri MP, Keerthana NS, Madhura M, Sankar A, Munnavar T (2018)Assistive technology for the visually impaired using computer vision. In: 2018 international conference on advances in computing, ICACCI. IEEE, pp 1725–1730
19. Kumar S, Mathew S, Anumula N, Chandra KS (2019) Portable camera-based assistive device for real-time text recognition on various products and speech using android for blind people. In: Innovations in electronics and communication engineering, pp 437–448
20. Mann WC, Hurren D, Karuza J, Bentley DW (1993) Needs of home-based older visually impaired persons for assistive devices. J Visual Impairment Blind 87(4):106–110
21. Elmannai W, Elleithy K (2017) Sensor-based assistive devices for visually-impaired people: current status, challenges, and future directions. J Sensors 17(3):565
22. Hakobyan L, Lumsden J, O'Sullivan D, Bartlett H (2013) Mobile assistive technologies for the visually impaired. J Sur Ophth 58(6):513–528
23. Min S, Oh Y (2020) Survey on obstacle detection features of smart technologies to help visually impaired people walk. J Korea Ind Info Sys Res 25(3):31–38
24. Massof RW Auditory assistive devices for the blind. Georgia Institute of Technology
25. Scheggi S, Talarico A, Prattichizzo D (2014) A remote guidance system for blind and visually impaired people via vibrotactile haptic feedback. In: 22nd mediterranean conference on control and automation. IEEE, pp 20–23
26. Palanisamy K, Arunkumar K, Bhuvaneshwaran P, Naveenkumar S, Dhamodharan M (2017) Walking stick with OPCFD system. Global Res Devel J Eng 3(1):1–5
27. Dhanuja R, Farhana F, Savitha G (2018) Smart blind stick using Arduino. J IRJET 5(03)

28. Aymaz Ş, Çavdar T (2016) Ultrasonic assistive headset for visually impaired people. In: 39th international conference on telecommunications and signal processing. IEEE, pp 388–391
29. Herghelegiu P, Burlacu A, Caraiman S (2017) Negative obstacle detection for wearable assistive devices for visually impaired. In: 2017 21st international conference on system theory, control and computing (ICSTCC). IEEE, Sinaia, pp 564–570
30. Kamal MM, Bayazid AI, Sadi MS, Islam MM, Hasan N (2017) Towards developing walking assistants for the visually impaired people. In: 2017 IEEE region 10 humanitarian technology conference (R10-HTC). IEEE, pp 238–241
31. Velázquez R (2010) Wearable assistive devices for the blind. In: Lay-Ekuakille A, Mukhopadhyay SC (eds) Wearable and autonomous biomedical devices and systems for smart environment. Lecture notes in electrical engineering, vol 75. Springer, Heidelberg.
32. Priya T, Sravya KS, Umamaheswari S (2020) Machine-learning-based device for visually impaired person. Artificial intelligence and evolutionary computations in engineering systems, pp 79–88
33. Dahlin, Ivanoff S, Sonn U (2004) Use of assistive devices in daily activities among 85-year olds living at home focusing especially on the visually impaired. J Disability Rehab 26(24):1423–1430
34. Panchanathan S, Black J, Rush M, Iyer V (2003) iCare-a user centric approach to the development of assistive devices for the blind and visually impaired. In: 15th IEEE international conference on tools with artificial intelligence. IEEE, pp 641–648
35. Krishna S, Colbry D, Black J, Balasubramanian V, Panchanathan S (2008) A systematic requirements analysis and development of an assistive device to enhance the social interaction of people who are blind or visually impaired. In: Workshop on computer vision applications for the visually impaired, James Coughlan and Roberto Manduchi, Marseille, France
36. Lan W, Dang J, Wang Y, Wang S (2018) Pedestrian detection based on YOLO network model. In: IEEE international conference on mechatronics and automation (ICMA). IEEE, pp 1547–1551
37. Shafiee MJ, Chywl B, Li F, Wong A (2017) Fast YOLO: a fast you only look once system for real-time embedded object detection in video. arXiv:1709.05943
38. Huang R, Pedoeem J, Chen C (2018) YOLO-LITE: a real-time object detection algorithm optimized for non-GPU computers. In: IEEE international conference on big data. IEEE, pp 2503–2510
39. Jeong HJ, Park KS, Ha YG (2018) Image preprocessing for efficient training of YOLO deep learning networks. In: IEEE international conference on big data and smart computing. IEEE, pp 635–637
40. Du J (2018) Understanding of object detection based on CNN family and YOLO. J Phys Conf Ser 1004(1):012029. IOP Publishing

A Performance Study of Prediction Models for Diabetes Prediction Using Machine Learning

Priya Mohan and Ilango Paramasivam

Abstract During the past couple of decades, the geriatric community is commonly affected by a well-known disease, namely diabetes. Recently, many researchers are focusing on developing a prediction model which can accurately predict if the patient is affected by diabetes at an early stage so that they can prevent further complications in health. The proposed research work focuses on analyzing the performance of various machine learning algorithms such as logistic regression, support vector machine, KNN, random forest, naïve Bayes, and gradient boosting classifier which could be used as a prediction model for predicting the common disease diabetes. The performance of these machine learning algorithms is compared, evaluated, and validated using the accuracy score. The results show that random forest classifier outperforms all the other classification algorithms considered for the study.

Keywords Prediction model · Machine learning · Classifiers · Performance · Evaluation · Health care

1 Introduction

In recent few couples of decades, there are various diseases by which most of the geriatric community are affected like Alzheimer's, Parkinson's, and diabetes. Out of these, diabetes is the basic disease that gives rise to other major diseases, and also, the elderly is at risk of being affected by stroke, liver damage, etc., which might lead to their death. As per the International Diabetes Federation, approximately 451 million people are affected by diabetes during the year 2017. The federation expects that the current count might increase to 693 million people during the fore coming

P. Mohan (✉)
Department of Computer Science, Bharathiar University, Coimbatore, Tamilnadu 641046, India
e-mail: priya.vinoth13@gmail.com

I. Paramasivam
Department of Computer Science & Engineering, PSG Institute of Technology and Applied Research, Coimbatore 641062, India
e-mail: dr.p.ilango@gmail.com

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_4

years [1]. Diabetes is a kind of chronic disease in which the glucose level of blood in the human body is inconsistent since the pancreas does not function properly due to which the insulin secretion level is disturbed and leads to diabetes. This is considered a chronic disease as there is no cure for this and can only be managed by medication. The elderly people who are affected by diabetes are at risk of being affected by heart disease, nerve damage, liver damage, etc. Thus when this disease is predicted at an early stage, it can be treated at the right time which will prevent any further complications.

Data mining is the discovery of knowledge and useful information from the large amounts of data stored in databases. It is referred to as knowledge discovery from databases (KDD), which is the automated or convenient extraction of patterns representing knowledge implicitly stored in large databases. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Even for security attacks, data mining techniques are applied for detection [2]. Data mining techniques have been widely used in diagnostic and healthcare applications [3, 4].

2 Literature Survey

Kemal Polat et al. presented a novel method for the diagnosis of hepatitis disease. Their method is based on a hybrid method that uses feature selection and an artificial immune recognition system (AIRS) with a fuzzy resource allocation mechanism. The AIRS has shown an effective performance on several problems such as machine learning benchmark problems and medical classification problems like breast cancer, diabetes, and liver disorders. By hybridizing FS and AIRS with a fuzzy resource allocation mechanism, a method is obtained to solve this diagnosis problem through classification. The robustness of this method concerning sampling variations is examined using a cross-validation method, and the dataset is taken from the UCI machine learning repository. They obtained classification accuracy, sensitivity, and specificity values for the diabetes disease dataset.

A. H. Roslina reported that hepatitis patients are those who require uninterrupted particular medical treatment to decrease and the death rate. By scientific investigation findings, data and machine learning techniques such as support vector machine (SVM), the classification and prophecy of their life diagnosis can be done. Even if, it cannot promise with the aim of all the features values in the data are related to one another. Thus, they integrate wrapper methods to eradicate noise features before classification. Their study shows the enhancement of prediction among data by connecting feature selection method earlier to classification method.

Shashikant Ghumbre et al. reported that medicinal analysis is measured as art against all equivalence intention made, which is significant because medicinal analysis requires the ability in coping with uncertainty simply not found in today's computing machinery. Using machine learning techniques, build up the software to help doctors in making an assessment exclusive of necessitating thorough discussion with the specialists. In their work, the application of artificial intelligence in typical heart disease identification has been investigated. In this study, an intelligent method-based support vector machine along with a radial basis function network was presented for the diagnosis. They developed an expert system based on clinical symptoms that are used to decide what type of heart disease is probable to come out of a patient and whether it is a heart attack or not. The sequential minimal optimization algorithm is applied to India-based patients' dataset, and then, the radial basis function (RBF) network structure trained by orthogonal least square (OLS) algorithm is applied to the same dataset for predictions. The results obtained support vector machine that can be successfully used for diagnosing heart disease [8].

Sahar A. Mokhtar et al. presented breast cancer screening that is the most effective and available tool. Nevertheless, the little up predictive rate of breast biopsy ensuing from mammogram interpretation leads to approximately 70% of redundant biopsies with benign outcomes. The data mining algorithms might be used to assist physicians in their decisions to make a breast biopsy on a doubtful lesion seen in a mammogram image or to do a short-term follow-up examination in its place. Using data mining classification algorithms, artificial neural network, decision tree, and support vector machine are investigated on mammographic masses dataset. Their idea of this study was to enlarge the physician's capacity to resolve the severity (benign or malignant) of a mammographic mass lesion from BI-RADS attributes and the patient's age. The dataset is separated for training the models and testing them by the ratio of 70%:30%, respectively, and the performance of classification algorithms is related through sensitivity, specificity, and classification accuracy. The accuracy of decision trees, artificial neural networks, and support vector machines are 78.12%, 80.56%, and 81.25% of test samples, respectively. Their result shows that the SVM classification model predicts the severity of breast cancer with the smallest amount of error rate and maximum accuracy [6].

K. C. Tan et al. used their work and real-life datasets are often interspersed with noise, making the subsequent data mining process difficult. The task of the classifier could be simplified by eliminating attributes that are deemed to be redundant for classification, as the retention of only pertinent attributes would reduce the size of the dataset and subsequently allow a more comprehensible analysis of the extracted patterns or rules.

3 Classification Algorithms Used in this Work

Classification is an important task in machine learning and data mining, which aims to classify each instance in the data into different groups. The feature space of a classification problem is a key factor influencing the performance of a classification learning algorithm [5].

The classification algorithms used in this research paper are

- Support vector machine (SVM)
- Naïve Bayes (NB)
- Random forest (RF)
- Logistic regression (LR)
- K-nearest neighbor (KNN)
- Gradient boosting classifier (GBC).

3.1 Support Vector Machine (SVM)

Support vector machine (SVM) is one of the classification algorithms that is used in this research paper. Vapnik and Chervonenkis in the year 1992 discovered SVM as a supervised learning technique that acts as a classifier between attributes. Hyperplanes are formed between attributes, and the main purpose of SVM is to maximize the margin between the hyperplane and the nearest points of every class. Support vector machine (SVM). It is mainly used for classification. SVM works on the principle of margin calculation. It basically draws margins between the classes. SVM has strong regularization properties. Regularization refers to the generalization of the model to new data. Support vector machines in particular were designed as a tool to solve supervised learning classification problems [9].

3.2 Naïve Bayes (NB)

The naïve Bayes algorithm is rooted in Bayes' theorem, a model developed by Thomas Bayes. Naïve Bayes considers that every feature is independent or does not have any correlation with other features. The existence or absence of one feature will not affect the other feature and is called Naïve. However, Bayes is a conditional probability that predicts the outcome of a future event based on the current condition. Naïve Bayes or also known as a probabilistic classifier is a fast, easy, powerful algorithm that can handle a large dataset. By using a training dataset collected from the data collection part, a model of the naïve Bayes algorithm is generated [7]. Furthermore, the mathematical equation is very easy to understand which are as follows:

$$P\,(y\,|z) = (P(zy)\,P(y))/(P(z)) \tag{1}$$

Based on the above equation:

$P(y)$ = Probability of event y occurring (hypothesis).

$P(z)$ = Probability of event z occurring (outcome).

$P(y|z)$ = Probability of event y occurring prior to event z is true.

$P(z|y)$ = Probability of event z occurring prior to event y is true.

3.3 Random Forest (RF)

Another classification algorithm that is used in this research paper is the random forest algorithm. It can only be used in supervised machine learning which can do both classification and regression activities. In this method, the random forest will do its work by generating a forest with decision trees [10]. Same as other classification methods, the more decision trees it generates, the higher accuracy it will achieve. The random forest algorithm is very similar to the ensemble approach. It used the principle of combining multiple "weak learners" to become a "strong learner" at the end. Random forest is a popular method of machine learning.

3.4 Logistic Regression (LR)

Logistic regression is another technique borrowed by machine learning from the field of statistics. It is the go-to method for binary classification problems (problems with two class values). Logistic regression is like linear regression in that the goal is to find the values for the coefficients that weight each input variable. Unlike linear regression, the prediction for the output is transformed using a nonlinear function called the logistic function. Logistic regression is used for classification, not regression. Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc., but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic Regression Mathematical Equation

Let the linear regression equation of y on x be

$$y = a + bx$$

When $x = xi$, the observed value of y is yi and the predicted value of y is $a + b_{xi}$. So,

$$ei = yi - (a + b_{xi}) \tag{2}$$

This ei is the error in taking $a + b_{xi}$ for yi. This is called the error of estimation. The method of least squares requires that a and b be so determined that

$$\sum_i (yi - a - bx_i) = 0 \text{ or } \sum_i y_i = na + b\sum_i x_i \tag{3}$$

$$\text{And} \sum_i (y_i - a - bx_i)x_i = 0 \text{ or}, \quad \sum_i xi\ yi = a\sum_i xi + b\sum_i x_i^2 \tag{4}$$

3.5 K-Nearest Neighbors (KNN)

KNN is a nonparametric method used for classification. It is also one of the best-known classification algorithms. The principle is that known data are arranged in a space defined by the selected features. When new data is supplied to the algorithm, the algorithm will compare the classes of the k closest data to determine the class of the new data. The major advantage of the KNN classification is its simplicity, it is also an efficient method. However, despite its efficiency, computation times can be long with large databases, the determination of the number of neighbors to use (k) requires trial and error, and the algorithm is weak with outliers which can strongly impact its efficiency.

K-nearest neighbors look at the k observations closest to a test observation x_0 and estimate the conditional probability that it belongs to class j using the formula

$$Pr(Y = j|X = x_0) = 1k\sum i \in N_0 I(yi = j)\ldots. \tag{5}$$

where N_0 is the set of k-nearest observations and $I(yi = j)$ is an indicator variable that evaluates to 1 if a given observation (xi, yi) in N_0 is a member of class j, and 0 if otherwise. After estimating these probabilities, k-nearest neighbors assign the observation x_0 to the class in which the previous probability is the greatest.

3.6 Gradient Boosting Classifier (GBC)

Gradient boosting is a machine learning algorithm and used for both classification and regression problems. It works on the principle that many weak learners (e.g., shallow trees) can together make a more accurate predictor [11]. Advantages of gradient boosting are:

The gradient descent technique to get the values of these coefficients gamma (Y), such that we minimize the loss function. Now let us dive deeper into this equation and understand the role of the loss function and gamma

$$F_{n+1}(X) = F_n(X) + \gamma n H(x, e_n) \tag{6}$$

$$L = \left(y - y'\right)^2$$

$$L = (y - (Fp(x))^2$$

Now, we know the error in our equation of $F_{n+1}(X)$ is the actual value minus updated predictions from all the models.

$$\frac{\mathrm{d}L}{\mathrm{d}\gamma} = (y_i - \gamma_i) = -(\text{observed} - \text{predicted})$$

The formula of residuals above, we see that the derivative of the loss function is multiplied by a negative sign, so now we get

$$(\text{Observed} - \text{Prediction})$$

The predicted value here is the prediction made by the previous model. In our example, the prediction was made by the previous model.

4 Relative Work

In the healthcare domain, machine learning algorithms are widely used to predict the occurrence of a disease at an early stage. The researchers had tried to use a variety of classifiers to predict the diseases and have obtained good accuracy results. They classified and analyzed the performance using the universally accepted dataset from the UCI repository. The results were evaluated using the parameters like accuracy, sensitivity, and specificity. They performed the classification in two different cases, one with preprocessed data and the other without preprocessing.

5 Framework for the Proposed Comparative Performance Study

5.1 Proposed Experimental Setup

The experimental setup and flow of the prediction model analysis are discussed in Fig. 1. The steps involved in the framework are discussed below:

Step1: Dataset is created from a widely well know Pima Indians Diabetes dataset. It encompasses female patients' details and their features.

Step 2: The first step of data preprocessing is missing value treatments. Check if there are any missing values in the data. Fill the missing values using attribute mean.

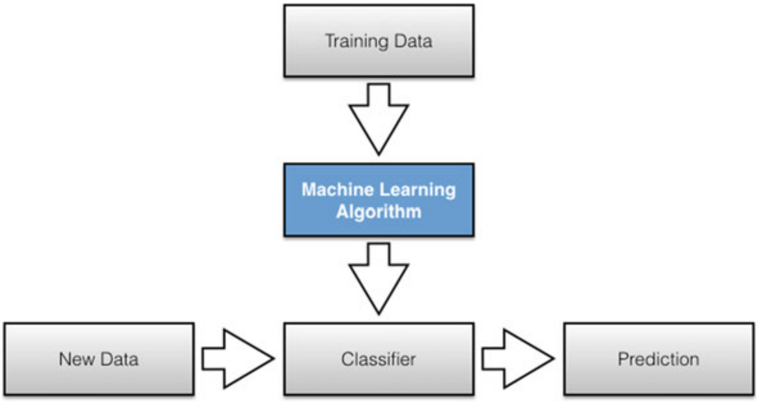


Fig. 1 Prediction model analysis

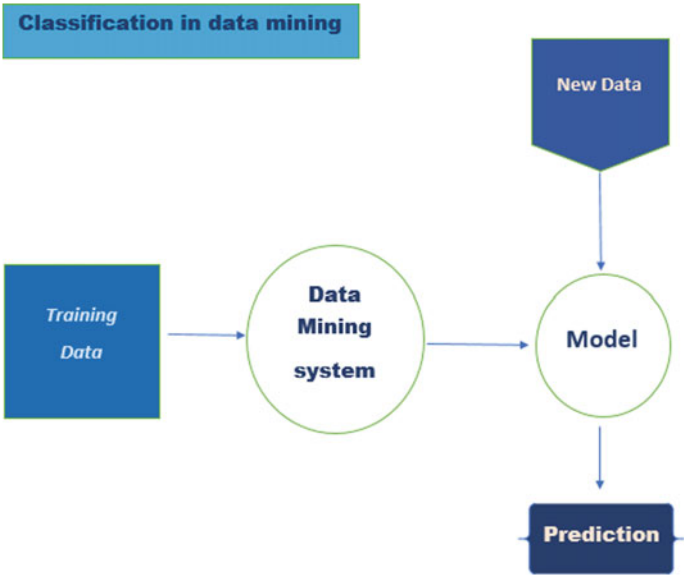


Fig. 2 Data mining system with classification algorithms

Step 3: The preprocessed data are predicted using logistic regression (LR), support vector machine (SVM), K-nearest neighbor classifier (KNN), random forest classifier (RF), naïve Bayes theorem (NB), and gradient boosting classifier (GBC).

Step 4: Evaluate the performance of the machine learning algorithm using standard measures such as accuracy score.

6 Dataset Description

The Pima Indians Diabetes dataset includes 768 complete instances described by 8 features (labeled as several times pregnant, glucose tolerance test, diastolic blood pressure, triceps skinfold thickness, 2-h serum insulin, body mass index, diabetes pedigree function, and age). The class distribution is class value 1 interpreted as "tested positive for diabetes" in 268 instances and class value 0 is interpreted as "tested negative for diabetes" in 500 instances.

The attribute evaluator is the technique by which each attribute in the dataset (also called a column or feature) is evaluated in the context of the output variable (e.g., the class). The search method is the technique by which to try or navigate different combinations of attributes in the dataset to arrive at a shortlist of chosen features.

Some attribute evaluator techniques require the use of specific search methods. For example, the CorrelationAttributeEval technique used in the next section can only be used with the ranker search method that evaluates each attribute and lists the results in a rank order. When selecting different attribute evaluators, calculate the correlation between each attribute and the output variable and select only those attributes that have a moderate-to-high positive or negative correlation (close to -1 or 1), and drop those attributes with a low correlation (value close to zero).

Running this on our Pima Indians Diabetes dataset suggests that one trait (plasma glucose) has the highest correlation with the output class. It also suggests a host of attributes with some modest correlation (mass, age, pregnancy). If we use 0.2 as our cutoff for relevant attributes, then the remaining attributes could be removed (diabetes pedigree function, insulin, skin thickness, and blood pressure).

Table 1 Dataset description

Column name	Data description	Unit measured	Range
Pregnancies	Number of times the female had been pregnant	Numeric	0–17
Plasma glucose	The concentration of the plasma glucose taken by glucose tolerance test	Numeric	0–199
Blood pressure	The diastolic measure of blood pressure	mm/Hg	0–122
Skin thickness	The thickness of skin fold	mm	0–110
Insulin	The measure of insulin in 2 h	muu/ml	0–744
BMI	Body mass index of the female	Kg/height in m2	0–80.6
Diabetes pedigree function	Functional measurement	Numeric	0.078–2.42
Age	Age of the female	Date	21–81
Outcome	A normalized value between 0 and 1 0: NO diabetes 1: affected by diabetes	Boolean	0 or 1

Feature selection via information gained using the InfoGainAttributeEval Attribute Evaluator. Like the correlation technique above, the Ranker Search Method must be used on our Pima Indians Diabetes dataset we can see that one attribute contributes more information than all of the others (Plasma glucose). If we use an arbitrary cutoff of 0.05, then we would also select the mass, age, and insulin attributes and drop the rest from our dataset.

6.1 Data Characteristics

To verify the categorical values and their distributions, we plotted these values in a histogram which is plotted in Fig. 3. This plot helped us to decide the closeness of the categories.

Also, we analyzed the target class to check the distribution of 0 and 1 options that is the relation between the number of patients with and without diabetes.

The scatter plot generated for the entire dataset is shown in Fig. 4. The plot gives the correlation as well as the outliers in the dataset we visualize instances that were outliers and can be eliminated for further prediction. Hence, we removed them. When these outliers were removed, more than 80 records were eliminated.

Performance evaluation

The performance of widely used prediction techniques is analyzed over the preprocessed dataset. The models which are used for predicting diabetes disease are

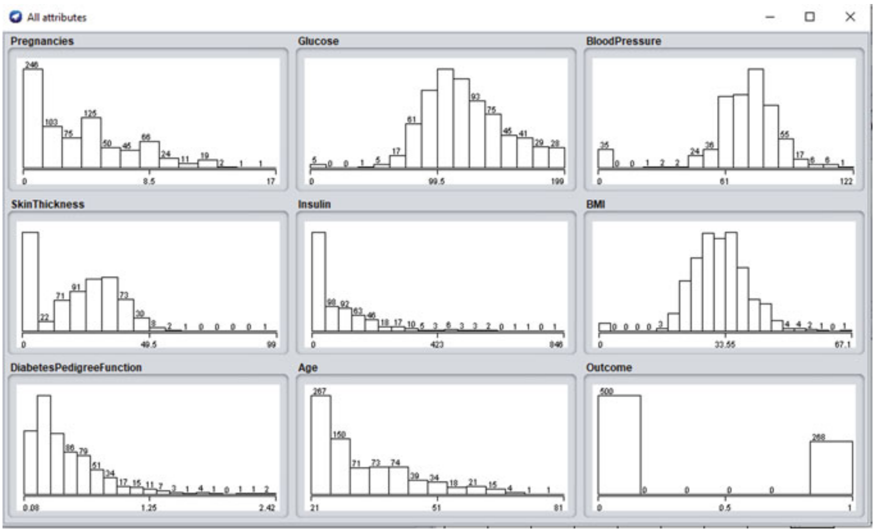


Fig. 3 Slot allocation

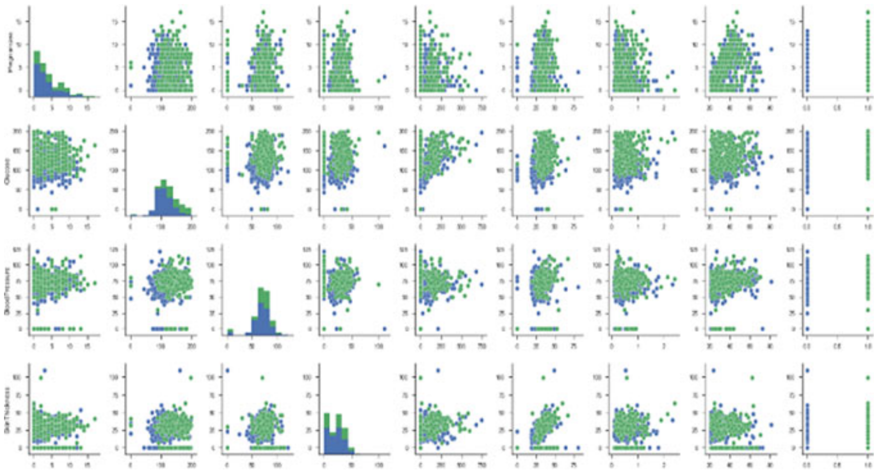


Fig. 4 Scatter plot for actual dataset

Table 2 Comparative analysis of performance for prediction models

Prediction model	Accuracy score
Logistic regression (LR)	0.7945
Support vector machine (SVM)	0.7975
K-nearest neighbor (KNN)	0.8611
Random forest (RF)	0.9667
Naive Bayes (NB)	0.7975
Gradient boosting classifier (GBC)	0.8761

logistic regression (LR), support vector machine (SVM), K-nearest neighbor classifier (KNN), random forest classifier (RF), naïve Bayes theorem (NB), and gradient boosting classifier (GBC). Accuracy is used for analyzing the performance of classification algorithms. The accuracy obtained is tabulated in Table 2.

The performance comparison of the accuracy among all the prediction models is plotted in Fig. 5.

The performance of the classification algorithms on the Pima Indians Diabetes dataset. The tabulated results and the plotted graph very clearly show that the random forest prediction model outperforms all the other prediction models.

7 Conclusion

This research paper focuses on predicting the common disease in elderly people using the various widely used machine learning techniques in the current scenario, namely random forest classifier, gradient boosting, KNN, logistic regression, naive

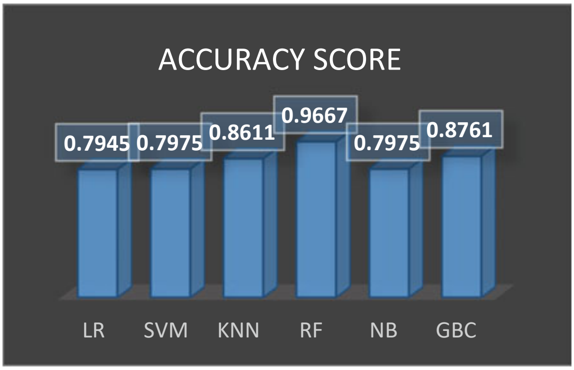


Fig. 5 Prediction model accuracy score

Bayes theorem, and SVM. The performance of these techniques is analyzed using the dataset which is available online, namely Pima Indians Diabetes dataset. The performance of these algorithms is compared based on the results obtained by taking 80% as training and 20% as testing data. The predicted results are already discussed in the previous section. The accuracy score of the algorithms is tabulated, and it clearly shows that the random forest classifier has the highest accuracy of 96.67% among all the others.

References

1. Cho N, Shaw J, Karuranga S, Huang Y, Fernandes JDR, Ohlrogge A, Malanda B (2018) IDF diabetes Atlas: global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes Res Clin Pr 138:271–281
2. Katiyar P, Senthil Kumarn U, Balakrishanan S (2013) Detection and discrimination of DDoS attacks from flash crowd using entropy variations. Int J Eng Technol 5(4):3514–3519
3. Prabha MS, Sarojini B (2019) Online healthcare information adoption assessment using text mining techniques. Mobile Netw Appl 24(4):1160–1165
4. Manonmani M, Balakrishnan S (2020) Feature selection using improved teaching learning based algorithm on chronic kidney disease dataset. Proc Comput Sci 171:1660–1669
5. Mokhtar SA, Elsayad AM (2013) Predicting the severity of breast masses with data mining methods. Int J Comput Sci Issues (IJCSI) March Issue 10(2)
6. Kandhasamy JP, Balamurali S (2015) Performance analysis of classifier models to predict diabetes mellitus. Procedia Comput Sci 47:45–51
7. Iyer A, Jeyalatha S, Sumbaly R (2015) Diagnosis of diabetes using classification mining techniques. Int J Data Min Knowl Manag Process 5:1–14
8. Polat K Güneş Salih(VL-88). A hybrid approach to medical decision support systems: Combining feature selection, fuzzy weighted pre-processing and AIRS. https://doi.org/10.1016/j.cmpb.2007.07.013
9. Palaniappan R, Sundaraj K, Sundaraj S (2014) A comparative study of the SVM and KNN machine learning algorithms for the diagnosis of respiratory pathologies using pulmonary acoustic signals. BMC Bioinform 15:223. https://doi.org/10.1186/1471-2105-15-223

10. Chen, RC., Dewi, C., Huang, SW. et al.(2020) Selecting critical features for data classification based on machine learning methods. J Big Data 7, 52. https://doi.org/10.1186/s40537-020-00327-4
11. Gradient boosting machines, a tutorial. Natekin, AlexeyKnoll, AloisSP-21, VL-7. https://doi.org/10.3389/fnbot.2013.00021

Orthopantomogram (OPG) Image Analysis Using Bounding Box Algorithm

Romir Mathur, Gopal Sakarkar, Kamlesh Kalbande, Rishi Mathur, Hrugved Kolhe, and Harish Rathi

Abstract The practice of AI in dentistry can increase dentist precision and save time. It can also provide a second opinion and confidence to the patient [1]. The system proposed in this paper works on ML, cloud computing, and python, which are trending tech in modern times. It promises to provide better performance with accurate results due to faster R-CNN implemented bounding boxes and the scope of batching and processing multiple OPG images in the cloud, real-time learning of ML model, and delivery of diagnosis report securely by web, mobile, and desktop applications. Even patients can donate their OPG images without their personal information to elevate the mastery of the ML model. This paper contains a section on methodology, result analysis, and conclusion. The methodology section speaks about the working of the model from the input of the raw OPG images, then the processing of the images with help of faster RCNN along with the bounding box. The result analysis describes machine learning model and presents processed images at each step. The initial image is a raw OPG and the final image is annotated with

R. Mathur · G. Sakarkar · H. Kolhe
Department of Artificial Intelligence Engineering, G H Raisoni College of Engineering, Nagpur, India
e-mail: mathur_romir.ai@ghrce.raisoni.net

G. Sakarkar
e-mail: gopal.sakarkar@raisoni.net

H. Kolhe
e-mail: kolhe_hrugved.ai@ghrce.raisoni.net

K. Kalbande · R. Mathur (✉)
Department of Electronics and Telecommunication Engineering, G H Raisoni College of Engineering, Nagpur, India
e-mail: mathur_rishi@outlook.com

K. Kalbande
e-mail: kamlesh.kalbande@raisoni.net

H. Rathi
Advance Dental Hospital, Ramdaspeth, Nagpur, India

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_5

the bounding boxes to represent the defected teeth. The final section describes future scope of the use of this technology for further development.

Keywords Dentistry · Artificial intelligence · Machine learning · Bounding box algorithm · Faster R-CNN · SuperAnnotate · Python · Firebase

1 Introduction

The application of artificial intelligence (AI) has already made many revolutions in the medical field. This paper proposes an AI-enabled system for the field of dentistry to analyze OPG images. The system starts to operate by feeding an input OPG image to a machine learning (ML) model. The input can get taken through a computer hard drive or cloud, and the model uses an object-detection algorithm known as the bounding box implemented through faster R-CNN to generate a diagnosis report. The report comprises all the defects in the input OPG image in a classified manner. The proposed system utilizes Firebase Cloud Storage (developed by Google) as a cloud-hosted database to store these reports in a more secure, accessible, and organized way. A large dataset of OPG images and a subset of this dataset labeled through SuperAnnotate software by expert dentists get used for training, validating, and testing the ML model, whereas python scripting language realizes the entire system architecture. This paper consists of the methodology to execute the system with the results of the ML model, cloud-hosted database explained in detail with the help of diagrams, flowcharts, and other figures.

2 Literature Review and Comparative Study

The paper used the AI-driven tool and further two deep convolutional neural networks (CNNs) were combined which provided lower accuracy for segmenting for upper and lower molars. The method used by us is faster R-CNN which is more efficient in object-detection and is time-efficient [1]. Faster R-CNN is able to detect the object much faster if the images are annotated whereas using Mask R-CNN which is based on faster R-CNN with an additional module of segmentation increases the time consumed hence the use of faster R-CNN is more viable computationally [2]. The paper uses a simple CNN network along with model tuning and dental image contrast enhancement which in turn gives an accuracy of 90% on the other hand the use of faster R-CNN is more time-efficient and improves accuracy without further techniques used [3]. The paper demonstrates how the use of faster R-CNN increases the accuracy and identifies if the teeth are present or not. Our paper has extended the use of faster R-CNN to also classify the defective nature of the teeth which provides the formation of the bounding box on only defective teeth [4]. This paper describes the use of faster R-CNN to find the place of the teeth this approach was similar to the one used by the model demonstrated by us. The usage of cloud computing technology was also added which helped deliver the results to patients and doctors quickly and saved time.

3 Methodology

The block diagram got shown in Fig. 1 displays the basic working of the proposed system. It takes an OPG image as input and feeds it to an ML model. The model then generates a diagnosis report which gets uploaded to a cloud-hosted database. Each block in the block diagram gets discussed further.

3.1 OPG Image

An Orthopantomogram (OPG) is a panoramic or wide view x-ray of the lower face. It displays all the teeth of the upper and lower jaw on a single film [1]. It should be stored virtually in a computer hard drive or cloud to feed into the proposed system. As the x-ray is virtually stored, it is referred to as an OPG image throughout this paper.

3.2 Machine Learning Model

The training and working of the ML model, executed by python, shown as a flowchart in Fig. 2. The state labels (S1–S5) shown in it are discussed later in the result analysis section of this paper.

A large dataset of OPG images and a subset of this dataset labeled by expert dentists get used for training, validating, and testing the ML model. The labeling/annotation of the dataset gets done through SuperAnnotate. It is an image annotation automation software that allows a model to train on high-quality and accurate data. It supports creating annotations regardless of team and project size,

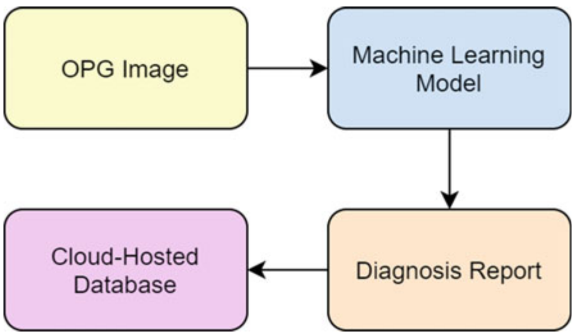


Fig. 1 Block diagram of the proposed system

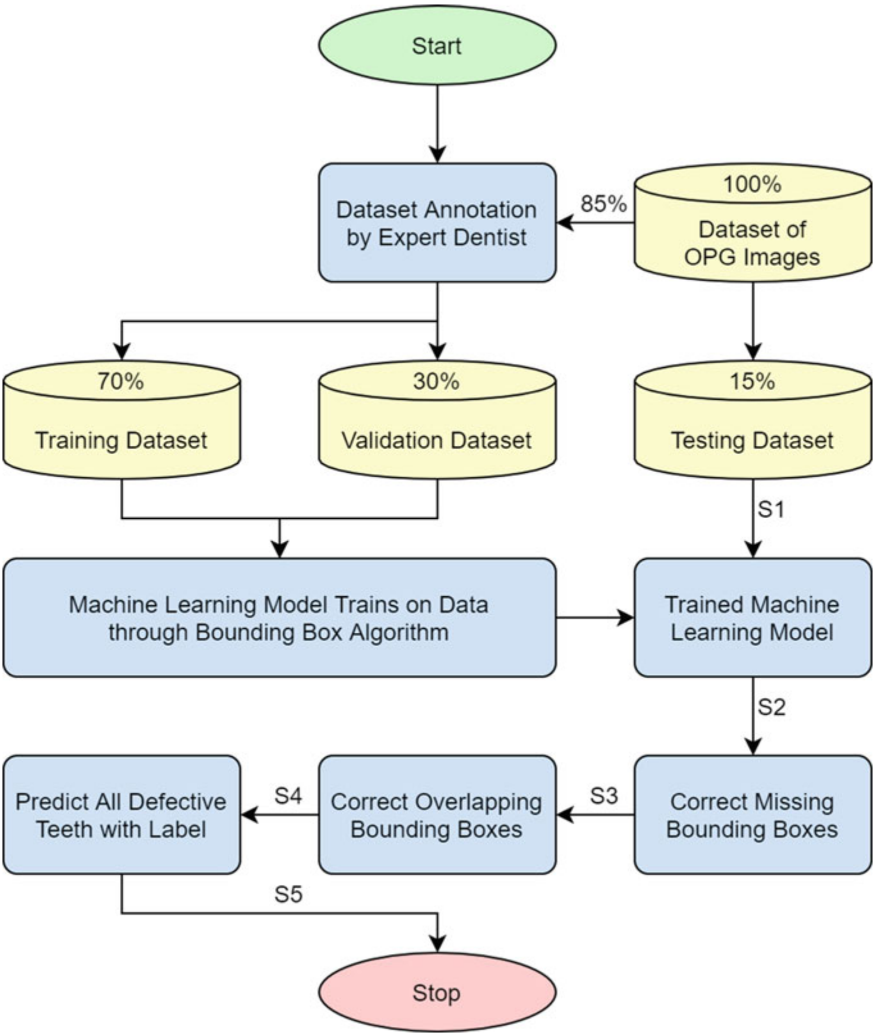


Fig. 2 Training and working of ML model

enables the use of neural networks to automate the annotation, and has multiple tools for image annotation like bounding box, polygon, ellipse, and more [2]. In the proposed system, steps used to annotate OPG images in the SuperAnnotate web platform get discussed further (Fig. 3).

The annotated data gets divided into training and validation datasets in ratio 7:3, and the ML model trains on both datasets through the bounding box algorithm. The bounding box is rectangular in shape and can get determined by the XY axis of the upper-left corner and lower-right corner of the rectangle [3]. It gets implemented through the family of R-CNN. For the proposed system, we have used faster R-CNN [4–7]. It provides better performance, computational efficiency, and reduction in test

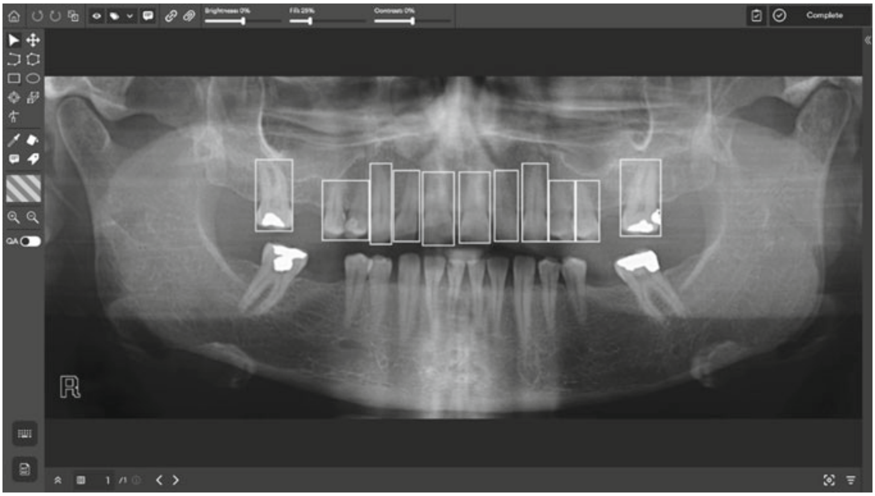


Fig. 3 The process of annotation of an OPG image

time. It consists of a regional proposal algorithm to generate the bounding boxes, a feature generation stage to obtain features from these bounding boxes, a classification layer to predict the class of the bounding box, and a regression layer to make the coordinates of the bounding box more precise [8]. Once the model gets trained, it gets fed with the testing dataset. The testing dataset is the remaining 15% of the original dataset consisting of unlabeled data. The trained model can apply bounding boxes on all teeth, correct missing bounding boxes, correct overlapping bounding boxes, and predict all defective teeth with a labeled class in the input OPG image.

3.3 Diagnosis Report

The final result of the ML model gets considered as a diagnosis report. It has the same format as the input OPG image. It contains bounding boxes applied on all teeth and all defective teeth classified with a label. For implementation purposes, the classes in the report will have an original name instead of a generic one. This report shows the number of teeth's defected out of all available tooth's. In Fig. 9, its show that five tooth's are defected along with their position in jaw.

3.4 Cloud-Hosted Database

The proposed system utilizes Firebase Cloud Storage as the cloud-hosted database. It is a powerful and cost-effective object storage service developed by Google. It gives robust operations, real-time updates, strong security, and high scalability and can get used with mobile, web, and server [9]. The implementation of the database, executed by Python, gets shown via a flowchart in Fig. 4. The state labels (S6–S9) are discussed later in the result analysis section of this paper. At the start, the patient's ID number gets taken as input. Then, the proposed system checks for a folder present in the database with that ID number. If the folder gets found, it stores the count of the number of files inside it. Otherwise, it creates a folder with that ID number in the database and initializes the count value equals zero. After which, it takes the diagnosis report as an input, increments the count value by one, and adds this updated value to the suffix of the report name before uploading it to the database [10, 11].

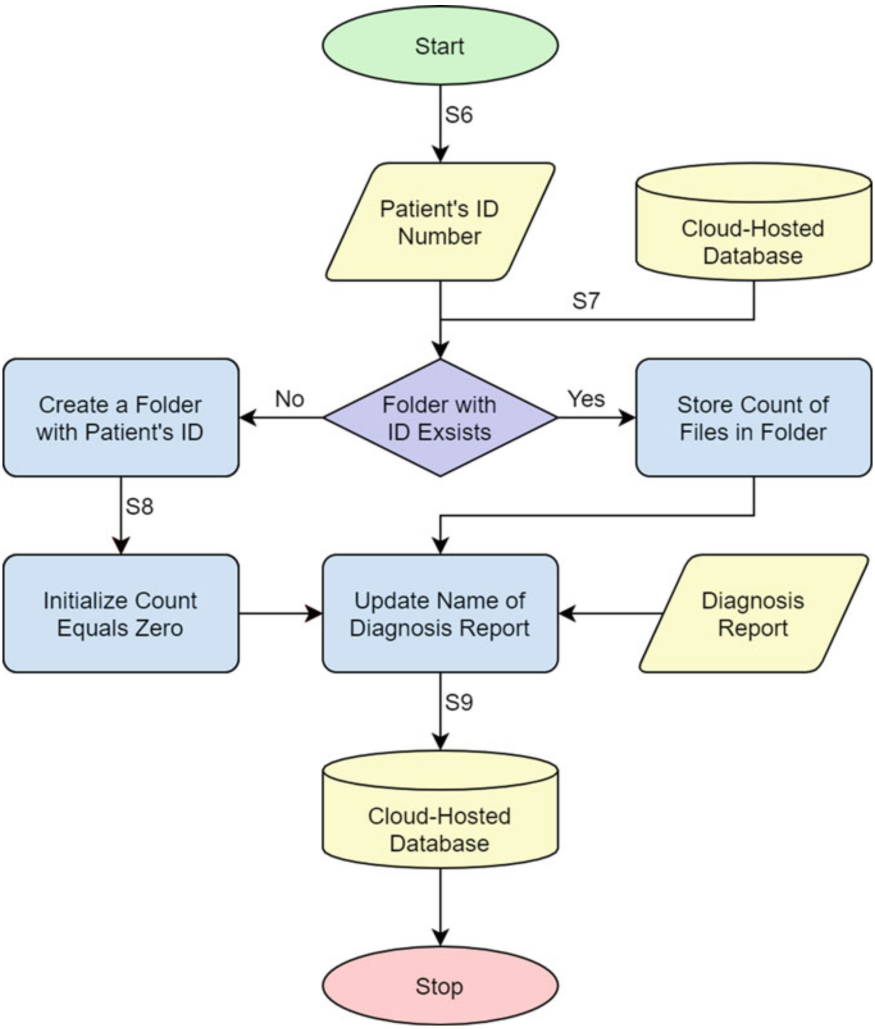


Fig. 4 Implementation of the cloud-hosted database

4 Result Analysis

Discussed further are the expected results of the ML model and cloud-hosted database in the proposed system based on each state label at which they got generated.

4.1 Machine Learning Model

At state S1 in Fig. 2, an OPG image gets selected at random from the testing dataset of the ML model, as shown in Fig. 5.

At state S2 in Fig. 2, the model starts processing the input OPG image and the first intermediate result records that it identifies most of the teeth but not all. It does not apply a bounding box on unidentified teeth. Even it implements overlapping bounding boxes on few distinguished teeth, as shown in Fig. 6.

At state S3 in Fig. 2, the second intermediate result records that the model has identified all the teeth and applied the bounding boxes to them but, a few of them are still overlapping, as shown in Fig. 7.

At state S4 in Fig. 2, the third intermediate result records that the model has identified all the teeth and applied non-overlapping bounding boxes to them. But it has not classified defective teeth yet, as shown in Fig. 8.

At state S5 in Fig. 2, the model finishes processing the input OPG image. The final result records that it has identified all the teeth. Also, it has applied non-overlapping bounding boxes on them and classified all the defective teeth, as shown in Fig. 9.

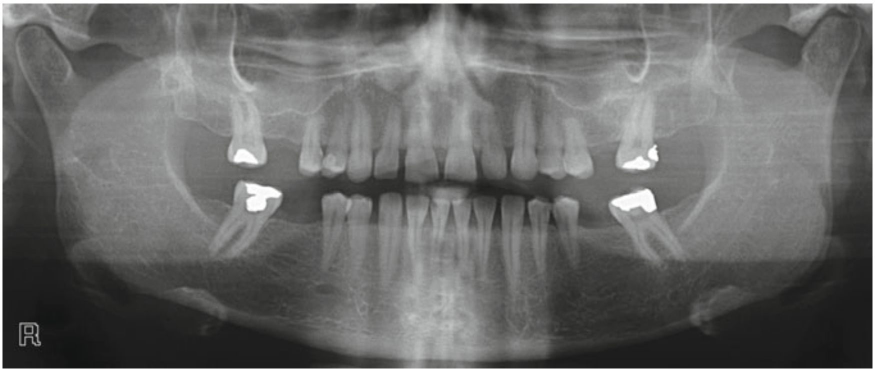


Fig. 5 An OPG image from the testing dataset

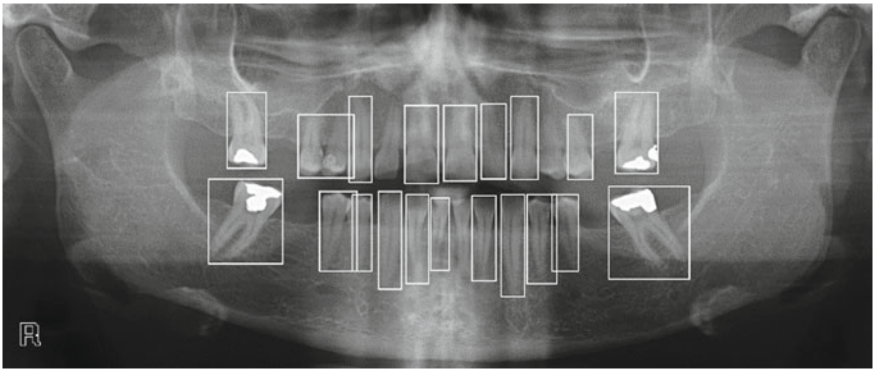


Fig. 6 The first intermediate result of the ML model

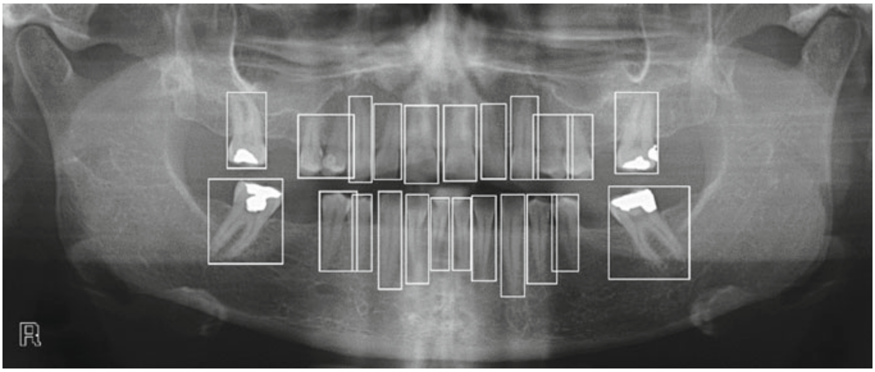


Fig. 7 The second intermediate result of the ML model

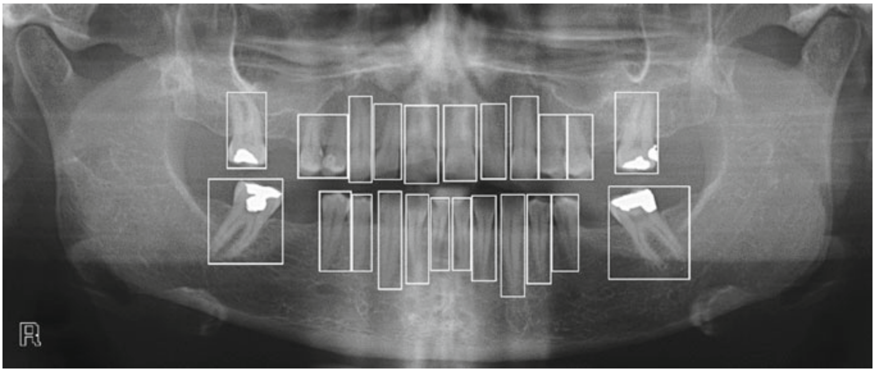


Fig. 8 The third intermediate result of the ML model

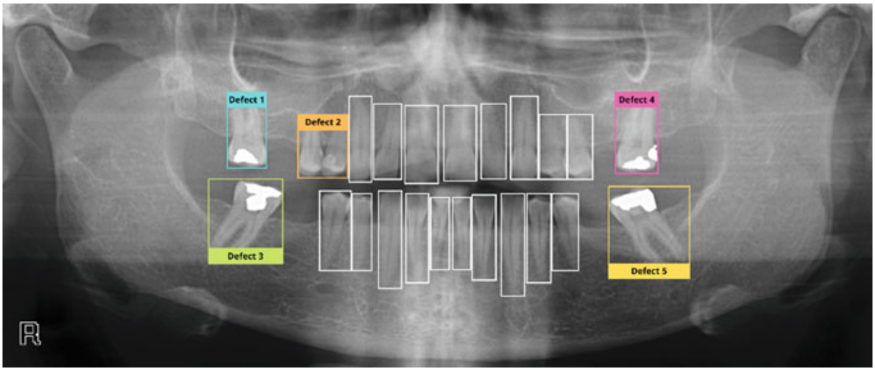


Fig. 9 The final result of the ML model

4.2 Cloud-Hosted Database

At state S6 in Fig. 4, to test, we enter a random ID number for the patient that is "002". State S7 in Fig. 4 shows the initial state of the cloud-hosted database, which has no folder named with this ID number, as shown in Fig. 10 (L). At state S8 in Fig. 4, the proposed system creates a new folder named with that missing ID number and initializes the count value equals zero, as shown in Fig. 10 (R).

After giving the final result of the ML model or diagnosis report as input to the system, it increments the count value by one and adds the updated count value to the suffix of the report name ("Diagnosis_Report_"). At state S9 in Fig. 4, it uploads the report to the database, as shown in Fig. 11.

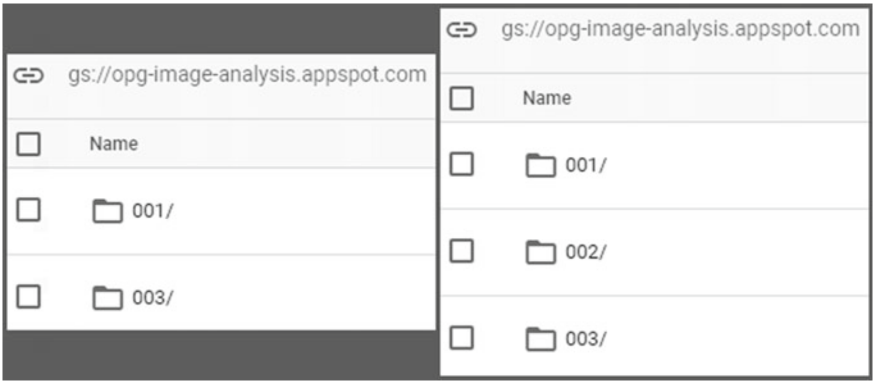


Fig. 10 The initial and processed state of the cloud-hosted database

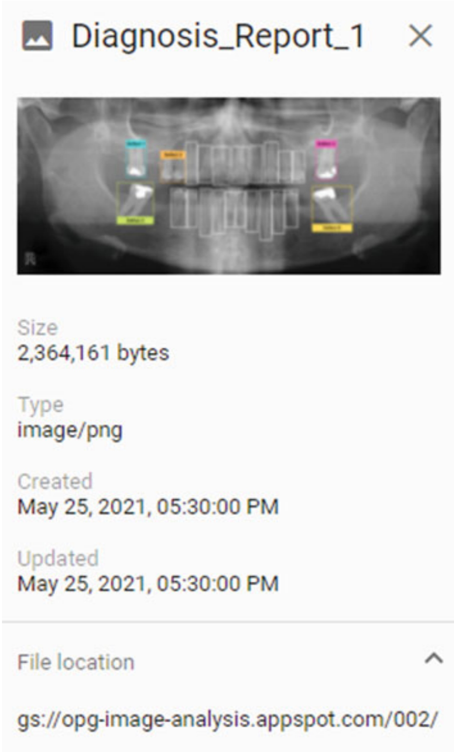


Fig. 11 Diagnosis report uploaded to the cloud-hosted database highlighting missing teeth's in color labels

5 Conclusion and Future Scope

Altogether AI and ML play a vital role in analyzing the OPG and discovering the deformities present. Machine algorithms such as CNN are very crucial for segmenting the OPG and discovering underlying properties for further analysis. These algorithms are very compatible to use and produce great results. R-CNN is a highly robust algorithm that helps in training the dataset that contains the bounding box. Tools such as SuperAnnotate help in annotating the OPG which gives a boost in training and validating the dataset. In the future, the same algorithm can be modulated and implemented in calculating the teeth density, detection of dental caries, measuring gaps in the teeth. A slight modulation of the algorithm can be very useful in the future and can be implemented by multiple clinics which will be easy to implement and cost-efficient for many patients.

References

1. I-MED Radiology Network OPG Procedure, https://i-med.com.au/procedures/dental-x-ray-opg. Last accessed 25 May 2021
2. SuperAnnotate Docs, https://doc.superannotate.com/docs. Last accessed 25 May 2021
3. Dive into Deep Learning Article on Bounding Box, https://d2l.ai/chapter_computer-vision/bounding-box.html. Last accessed 25 May 2021
4. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. IEEE Conf Comput Vision Pattern Recogn 2014:580–587
5. Girshick R (2015) Fast R-CNN. IEEE international conference on computer vision (ICCV) 2015:1440–1448
6. S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137–1149.
7. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask R-CNN. IEEE international conference on computer vision (ICCV) 2017:2980–2988
8. Towards Data Science Article on Faster R-CNN, https://towardsdatascience.com/faster-r-cnn-for-object-detection-a-technical-summary-474c5b857b46. Last accessed 25 May 2021
9. Firebase Documentation, https://firebase.google.com/docs. Last accessed 25 May 2021
10. Khanapurkar M, Kaushik R, Mathur R, Choudhari P, Gupta R, Dheer N (2021) Design and implementation of cloud-based smart parking management system. In: 2021 6th international conference for convergence in technology (I2CT), pp 1–4
11. Mathur R, Kalbande K (2020) Internet of things (IoT) based energy tracking and bill estimation system. In: 2020 fourth international conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), pp 80–85

Design and Analysis of an Improved Artificial Neural Network Controller for the Energy Efficiency Enhancement of Wind Power Plant

T. Mariprasath, C. Shilaja, CH. Hussaian Basha, M. Murali, Fini Fathima, and Shaik Aisha

Abstract Globally, Renewable Energy Resources (RER) are playing a vital role in generating the electrical energy due to the conventional fossil fuel-based power plants which are harming the environment. Also, the availability of fossil fuels is going to run out. The primary resources for RER are sun, wind, hydro, and tidal. Among energy, the harnessing rate has been rapidly increased in solar Photovoltaic (PV) and wind power plants. Since sun and wind energy are abundant in nature, nevertheless, natural resources are seasonal which are varying concerning the climatic condition. Therefore, sun and wind power generators are produced fluctuating electrical energy which causes stability issues. It can be compensated by the Maximum Power Point Tracking (MPPT) technique. At present, the MPPT technique is incorporated with RER for generating maximum electrical energy based on available resources. In this manuscript, a wind power plant with an Improved Variable Step-Radial Basis Functional Network (IVS-RBFN)-based MPPT model has been developed by using MATLAB/Simulink window to analyze the significance of MPPT. The simulation results show that wind power plants are capable of generating constant power with the help of IVS-RBFN-based MPPT technique. Furthermore, the wind power output is significantly enhanced with the accurately designed boost converter.

Keywords Boost converter · Convergence speed · Duty cycle · Enhancement of efficiency · Less steady state oscillations · Peak power point · Wind plant · Wide output voltage

---

T. Mariprasath · CH. Hussaian Basha (✉) · M. Murali · S. Aisha
K. S. R. M. College of Engineering (Autonomous), Kadapa, Andhra Pradesh, India
e-mail: sbasha238@gmail.com

M. Murali
e-mail: murali@ksrmce.ac.in

C. Shilaja
Academy of Research and Education, Kalasalingam University, Krishnankoil, Tamil Nadu, India

F. Fathima
Mar Baselios Christian College of Engineering & Technology, Kuttikkanam, Kerala, India

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_6

1 Introduction

Due to the rapid growth of urbanization and industrialization, the requirement for electrical energy has increased. Accordingly, power industries are enhancing the electrical energy generating capacity by capacity addition program. During capacity addition, fuels are playing a key role. Because of conventional thermal-based power plants, fossil fuels are used as a primary resource for generating thermal energy [1]. It is used to generate the required steam. The fossil fuels are namely coal, diesel, and petrol. Among, over 60% of power plants use coal as a primary fuel for generating required thermal energy [2]. The fossil fuel-based power plants faced two challenges such as the availability of fossil fuels, which is going to run out in the near further, and fossil-based power plants which are harming the environment. The coal-based power plants are releasing carbon dioxide, carbon monoxide, etc. This harms the environment as well as a living organism [3]. Therefore, power sectors are concentrating ecologically friendly resources for generating electrical energy. In addition, the availability is also considered for the generation of electrical energy.

Globally, renewable energy resources (RER) are playing a key role in extracting the electrical energy. RER are solar, wind, tidal, hydro, etc. Among them, electrical energy generation has been increased from solar and wind power plants. By using a suitable energy conversion device, it is possible to convert available RER into electrical energy [4]. For example, solar PV systems and solar PV arrays are used to convert available irradiations into electrical energy. The photovoltaic cells are connected in series which form the solar module. Then, several solar cell modules are connected to form a solar PV array. The basic principle behind the solar PV system is the photovoltaic effect. The PV cells received irradiations from the sun. The solar cell is made upon two different layers of silicon such as P and N-type semiconductor materials [5]. By nature, N-type semiconductor materials release the electron when sun irradiation is hit on the materials while P-type materials receive the extra electrons. It is the simplest principle behind the solar PV arrays.

The output of the solar PV array strongly depends on the environmental conditions such as available irradiations, irradiation received by the solar PV array, and cell temperature. The received irradiations by the solar PV cell are high and the generating electricity is also high whereas the low amount of irradiations is hit in the solar PV which yields the less energy. Therefore, solar PV array yield potential has been varied concerning the irradiations [6]. The sun continuously moves; therefore, received irradiation by the solar PV array has been varied concerning time. So, the electrical energy harnessing rate is varied concerning time. For example, the mid of the solar PV generate high energy, while morning and evening yield potential has been low than of mid the day. Therefore, extracting maximum energy is a challenging task that is achieved by keeping solar PV working on MPP [7]. It is made by the MPPT technique. The role of MPPT is to track the maximum power from available irradiations.

Traditional MPPTs such as Incremental Conductance (IC), Perturb and Observation (P&O), and Hill climbing can track the maximum power steady under climatic conditions. The wind power plant is not able to generate constant power since the out electrical energy obtained from the wind power plant strongly depends on the available wind speed. The wind speed depends on the atmosphere's conduction. Wind speed is continuously varying and, therefore, extracts maximum power from the crucial task [8]. Due to fluctuating output, these power generation units are interconnected with the utility grid which is a difficult issue. To rectify such a problem, the wind power plants are incorporated with MPPT. Traditionally, MPPT techniques are enhancing the energy yield potential of a solar PV system [9].

In P&O technique, at first, the voltage and current have been measured by using a suitable sensor. It is used to estimate the power. Next, instantaneous voltage and current have been measured. This is used to estimate instantaneous power. Whether the change in power is greater than zero or not has been checked. Whether the change in voltage is zero or not has been checked. At this moment, if the condition is true, then the duty cycle is increased while the duty cycle is getting reduced. Nevertheless, if change in power is greater than zero, then the duty cycle of the switching device is reduced. Otherwise, the duty cycle is increased [10]. Whenever a rapid climatic change occurs at instant soft computing-based MPPT technique is capable of tracking maximum power from the solar PV system. Here, in this work a Variable Step-Radial Basis Functional Network (VS-RBFN) is used for generating the switching pulses to the boost converter.

Furthermore, the electrical energy extraction rate is significantly increased from a wind power plant. The wind power plants are tracking electrical energy from wind. The wind is referred to the movement of air due to uneven heating of the earth's surface by the sun. From wind, electrical energy has been extracted by blades [11]. Globally, total installed capacity of the wind power plant is 300 GW which is rapidly increasing every hour. According to the International Energy Agency Report, wind power has been contributing 18% of power-sharing by 2050. However, wind power plants do not produce constant electrical energy since wind output energy depends on the available wind speed [12].

Suppose wind speed is high which produced a high energy yield otherwise vice versa. In addition, wind power yield potential has been varied concerning temperature, pressure, and humidity, respectively. Therefore, extracting maximum power from wind power plants is a challenging task. In this paper, a wind power plant with MPPT is modeled and analyzed with the help of a MATLAB/Simulink block set. Also, an accurate design has been developed for enhancing voltage from generation to the desired level. The manuscript has the following sections such as Sect. 2 deals with the proposed model block diagram description followed by Sect. 3 discussion about analysis of power plant. Section 4 discusses the design of VS-RBFN MPPT, Sect. 5 discusses simulation results and analysis, and Sect. 6 is concluding the main findings.

2 Proposed Model of Wind Power Plant Station

Fig. 1 shows the entire unit of the proposed model. The first block represents the wind turbine. It received three inputs such as generator speed (G_s), Pitch Angle (P_a), and Generator Base Speed (B_s), respectively. The turbine is generating the required kinetic energy for a permanent magnet synchronous generator. Here, salient pole synchronous generator has been used. Due to it, seed torque characteristics are comparable to the non-salient pole synchronous generator [13].

The PMSG is generating electrical energy depending on the received mechanical energy from the wind turbine. Subsequently, the generated electrical energy is fed to the bridge rectifier. It is used to rectify the given input power from AC to DC. It is fed to the input of the boost converter. The boost converter boosts the voltage at the desired level according to the inductor, capacitor, and switching sequence. The voltage level increases or decreases depending on the duty cycle of the switch. The duty cycle generation depends on the generator output voltage and rectifier voltage and power as shown in Fig. 1. The working converter inductor charging and discharging currents are given in Fig. 2. From Fig. 1, the conventional boost converter steps up the inverter output voltage by using the power point tracking controller. The switching behavior of the converter is given in Fig. 2.

From Fig. 2, the switch working time period is $D * T_s$, and its blocking time is represented as $(1 - D) * T_s$. When the switch is in forward mode, then the inductors start charging with the slope of V_{dc}/L. Similarly, if the switch is in ideal condition, then the currents of inductors start reducing with the slope of $(V_{\mathrm{dc}} - V_0)/L$. The voltage gain of the proposed converter is derived by using the following equations as,

$$DV_{\mathrm{dc}} * T_S + (1 - D) * (V_{\mathrm{dc}} - V_0) * T_S = 0 \tag{1}$$

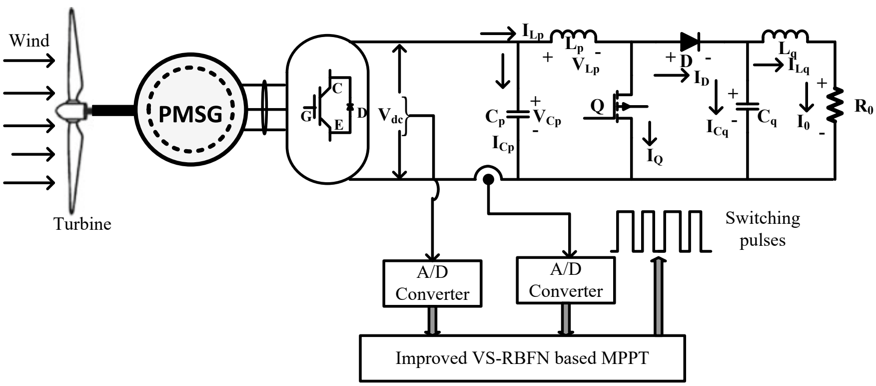


Fig. 1 Block diagram of proposed wind plant fed VS-RBFN-based MPPT controller

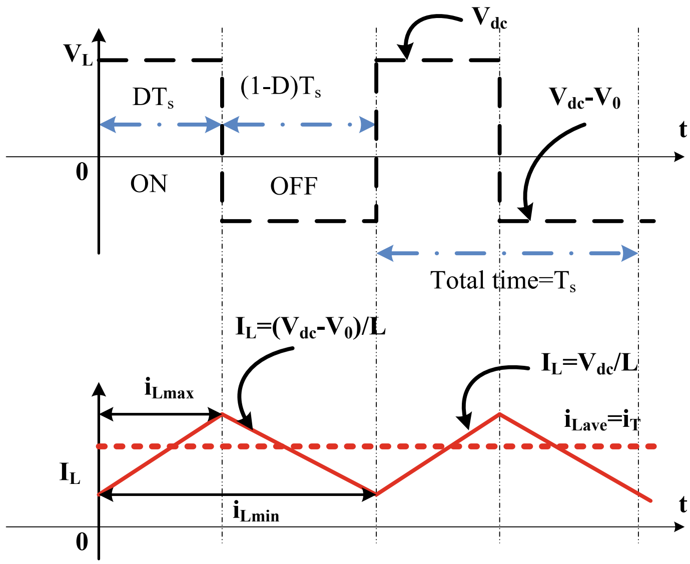


Fig. 2 Operating waveforms of conventional boost converter

$$-I_0 * DT_S + (1 - D) * (I_{\mathrm{dc}} - I_0) * T_S = 0 \quad (2)$$

$$V_0 = V_{\mathrm{dc}}/1 - D, \text{ and } I_0 = I_{\mathrm{dc}} * (1 - D) \quad (3)$$

$$V_0/I_0 = R_0, \text{ and } V_{\mathrm{dc}}/I_{\mathrm{dc}} = R_{\mathrm{dc}} \quad (4)$$

Based on the Eqs. (1)–(4), the terms D, V_{dc}, T_s, and V_0 are the boost converter duty cycle, voltage at input side, switch working time, and load voltage. Similarly, the terms R_0, I_0, and R_{dc} illustrated the converter load resistor, output current, and source equivalent resistance.

3 Design and Analysis of Wind Power Station

Traditionally, the squirrel cage induction motor is used to generate electrical energy which working based on the Danish concept, respectively. It is directly connected to the grid. The speed of such a motor is constant such as fixed speed [14]. Perhaps, at the instant of heavy wind hitting on the wind blades subsequently generating electrical energy is also high. Later on, the adjustable speed control technique has been incorporated with the existing squirrel cage induction motor. This restricts high wind pressure on wind towers [15]. In wind power plant, Double Filed Induction Generator (DFIG) is popularly used which is coupled to the turbine with help of a gearbox. Due to aging phenomena, the gear gets damaged. Therefore, the frequent maintenance is required for smooth mechanical coupling in between generator and the turbine. However, the rotor of PMSG is directly coupled to the wind turbine, so there is no need for a gear system.

So, cost-wise PMSG is lower than that of DFIG. The significance of PMSG is low excitation loss, reasonable efficiency, high power density, and low maintenance cost. However, the wind power plants introduced power quality issues on the transmission and distribution system. To compensate for the power quality issue at present dynamic voltage resistor (DVR), a thyristor-switched series capacitor (TSSC) has been incorporated with the existing system for enhancing the power quality. In addition, the voltage sag, and swell issues are reduced by using the DVR. Moreover, wind power plants cause following impacts on the power systems such as short and long duration effects [16]. The first effect time duration is marginally very low that is milliseconds to hours which is responsible for system unbalancing. When generated powers are transferred to the grid, it causes the power quality problem, voltage sag and swells, and so reactive power problems, respectively. Suppose the required compensating devices are not included which continues until wind power gets off.

Wind power plants never run at a constant speed. The main classifications of voltage variations are voltage sag and swell, and short and long voltage irruptions, respectively. It can be rectified by a suitable compensator such as STATCOM. It is a synchronous condenser that is connected parallel to the AC system, respectively. Nowadays, STATCOM is incorporated with a harmonics filter to attain effective control. On the other hand, power has been enhanced by SVS. The primary role of SVS is automatically matching the impedance with the required system. For example, if power system load is capacitive that of others, using reactor SVS consumes the VAR from the power systems. Nevertheless, the capacitor bank of SVS gets enabled when the power system is more inductive [17].

Due to the absence of excitation, PMSG is popularly used in wind power plants, so machine cost is low as well maintenance cost is low. Perhaps, it has low-speed characteristics. Therefore, the DVR system has been incorporated into the wind energy system for maintaining the system's consistency and reliability, respectively. The DVR has switching devices such as thyristor, MOSFET, or IGBT and voltage source inverter with a low pass filter [18]. The Pulse Width Modulation Technique has been used for triggering VSI. The role of the low pass filter is suppressing the harmonics content which is developed during voltage conversion from DC to AC, respectively. The modern power system has a separate control unit to match the voltage phase angle and frequency.

4 Design of Proposed Variable Step-RBFN Algorithm

From the previous existed articles, a neural network is a circuit consisting of number of neurons. In the neural network, each and every neuron acts as a node. Also, the network contains series of activity process in order to identify the basic relations

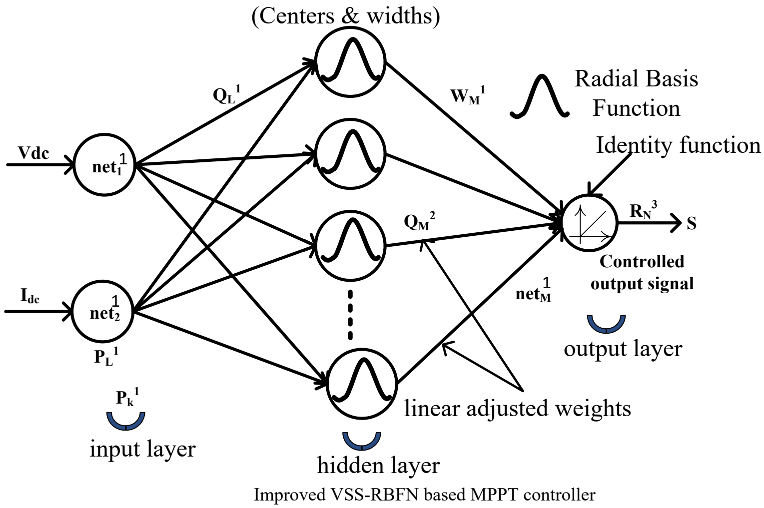


Fig. 3 Analysis and complete action of RBFN MPPT technique

in a set of data sets. The NN algorithms are used to identify the hidden patterns, correlations in raw data, text classification, semantic parsing, information extraction, paraphrase detection, multi-document summarization, and speech recognition. Here, in this work, a radial basis neural network is proposed because its features are good generalization, online learning ability, strong tolerance to input noise, and easy design. Also, this technique is useful to solve all the nonlinear complex problems. From Fig. 1, the signals considered for the operated MPPT controller are the wind-fed voltage and current. The complete action of radial basis neural network is given in Fig. 3.

The above showed neural network structure is used for obtaining of peak power position of the wind. It is one of the feed forward networks constituted of three layers and those are named as input later, middle layer with nonlinear RBF initiation, and linearized output layer. The input of the neural network can be modeled in the form of real vectors. The output of network is a scalar function of corresponding input. The mathematical derivations of RBF network are given as,

$$Q_L^1(t) = f_L^1(net_L^1(t)) = net_l^1(t); t = 1, 2, 3..n \tag{5}$$

$$net_M^2(t) = -(x - m_t)^T \sum\nolimits_t (i - m_t); t = 1.2.3.., n \tag{6}$$

From Eqs. (5) and (6), the variables Q, t, m, and net are expressed as hidden layer, number of iterations, mean value, and net value. Similarly, from Eqs. (7)–(10), the parameters *R*, *W*, *K*, and *e* are defined as output layer, weight of neuron, number of neurons, and error of the signal.

$$R_M^2(t) = f_M^2(\text{net}_M^2(t)) = \text{net}_M^2(t); t = 1, 2, 3 \ldots, n \tag{7}$$

$$net_N^3 = \sum\nolimits_M W_M * y_M^2(k); t = 1.2.3.., n \tag{8}$$

$$R_N^3(t) = f_N^3(net_N^3(t)) = net_N^3(t); t = 1.2.3..., n \tag{9}$$

$$e = \sum_{t=1}^{n} \frac{1}{2}(V_{\text{ref}} - V_{MPP}) \tag{10}$$

5 Simulation Results and Analysis

Figure 4 showed relation between turbine speed and turbine output power, and the maximum power has been delivered at a wind speed of 12 m/s. Speed and the proposed wind model have been developed in the MATLAB/Simulink to analyze the significance of the proposed IVS-RBFN MPPT. Figure 5 shows the relationship of the three-phase current of PMSG. From Fig. 5, it observes that each phase current has a 90 degree phase shift with each and other. Figure 6 showed the relation between mechanical characteristics of PMSG, such as rotor speed and electromagnetic torque. At initial, the electromagnetic torque and rotor speed are fluctuates continuously. This infers that the motor attains smooth speed very soon. Figure 7 shows the boost

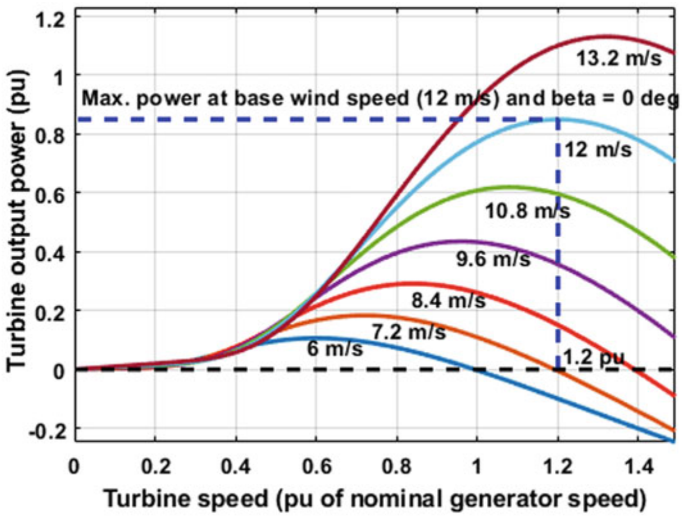


Fig. 4 Turbine speed versus turbine output power

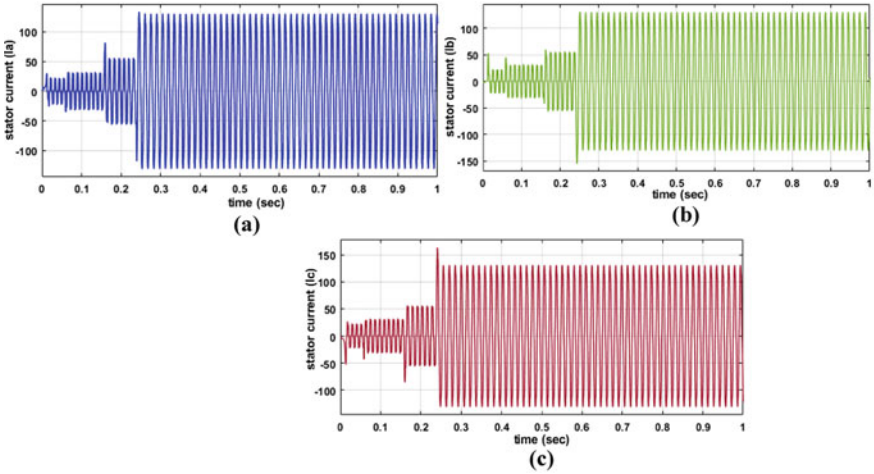


Fig. 5 Three phase stator current versus time

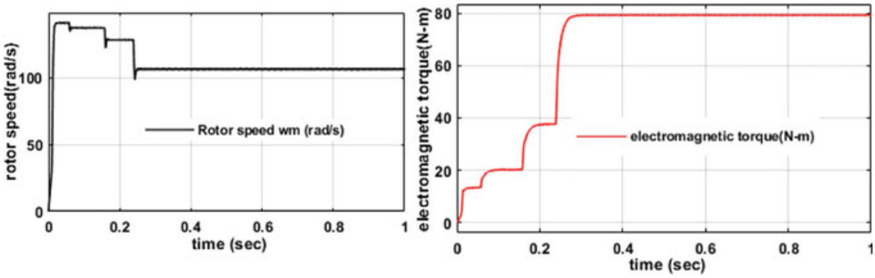


Fig. 6 Mechanical characteristics versus time

converter output voltage and current, respectively. The output of PMSG is almost perfect for three-phase voltage and current, respectively. Figure 8 observed that line voltage and current are in phase, and so at the peak of line voltage, a small fluctuation is present. This generated voltage is fed to the input of the universal bridge rectifier. The bridge rectifier is made by a non-gate device such as a diode.

6 Conclusion

The literature found that wind power plants not produce continuous and constant output. Therefore, it causes power quality issues on the power system. It is rectified by suitable compensating devices. Compared to all other compensating techniques, the hybrid compensating technique is costly but it suppresses all types of harmonics issues as well as maintains the system voltage at a desirable level. The wind power

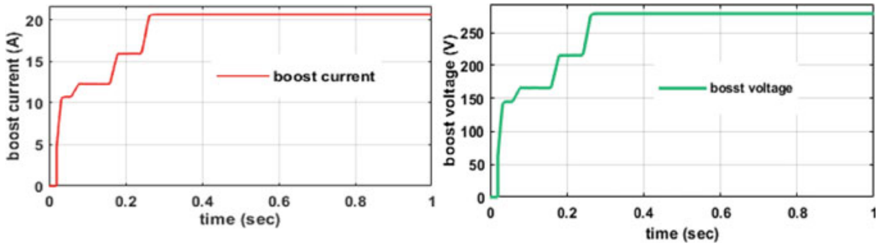


Fig. 7 Relation between boost converter output

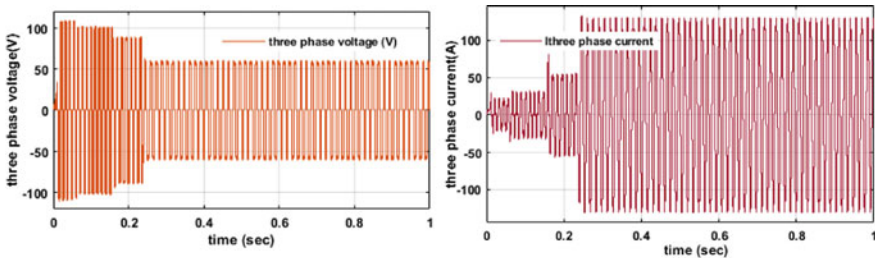


Fig. 8 Line voltage and line current relation

plant with MPPT is modeled and analyzed by using MATLAB/Simulink. The simulation found that, with IVSS-RBF MPPT, the developed model has produced constant power for the entire duration. Furthermore, with the help of a boost converter, step up the voltage which suppresses the transmission loss since wind farms are installed far away from the city center. In the future, the proposed MPPT technique is compared with other MPPT techniques to find effective MPPT for wind power plants.

References

1. Henry A, Ravi P, Arun M (2020) Five thermal energy grand challenges for decarbonisation. Nat Energy 5:635–637
2. Brauers H, Pao-Yu O (2020) The political economy of coal in Poland: drivers and barriers for a shift away from fossil fuels. Energy Policy 144:1–12
3. Hussaian Basha CH, Bhanutej JN, Rani C, Odofin S (2019) Design of an LPF based slider controller for THD reduction in solar PV B-4 inverter. In: International conference on electrical, computer and communication technologies. IEEE, pp 1–9
4. Hussaian Basha CH, Govinda Chowdary V, Rani C, Brisilla RM, Odofin S (2020) Design of SVPWM-based two-leg VSI for solar PV grid-connected systems. In: Soft computing for problem solving, vol 1048. Springer, pp 879–892
5. Govinda V, Chowdary, Udhay Sankar V, Derick M, Hussaian Basha CH, Rani C (2020) Hybrid fuzzy logic-based MPPT for wind energy conversion system. In: Soft computing for problem solving, vol 1057. Springer, pp 951–968

6. Corrado L, Andrigo Filippo A, Ricardo R (2017) The influence of different irradiation databases on the assessment of the return of capital invested in residential PV systems installed in different locations of the Brazilian territory. Sol Energy 155:893–901
7. Hussaian Basha CH, Rani C (2020) Performance analysis of MPPT techniques for dynamic irradiation condition of solar PV. Int J Fuzzy Syst 22:2577–2598
8. Jan S, Xiaorong X, Luping W, Wei L, Jingbo H, Hui L (2019) Overview of emerging subsynchronous oscillations in practical wind power systems. Renew Sustain Energy Rev 99:159–168
9. Bhattacharjee A, Samanta H, Banerjee N, Saha H (2018) Development and validation of a real time flow control integrated MPPT charger for solar PV applications of vanadium redox flow battery. Energy Convers Manage 171:1449–1462
10. Hussaian Basha CH, Rani C (2020) Design and analysis of transformer less, high step-up, boost DC-DC converter with an improved VSS-RBFA based MPPT controller. Int Trans Electric Energy Syst 30(12):1–21
11. Węgrzyński W, Lipecki T (2018) Wind and fire coupled modelling—part I: literature review. Fire Technol 54:1405–1442
12. Li C, Zhou S, Xiao Y et al (2017) Effects of inflow conditions on mountainous/urban wind environment simulation. Build Simul 10:573–588
13. Zhan Y, Kong K, Xu G, Kang J, Zhao H (2019) Analysis of damper transient currents in salient-pole synchronous generator with skewed armature slots considering interbar currents. IEEE Trans Ind Appl 55(1):336–343
14. Boldea I: (2017) Electric generators and motors: an overview. CES Trans Electrical Mach Syst 1(1):3–14
15. Zhaoxia L, Yong S, Jing Y, You Y, Hangfeng W (2017) Simulations and field-test of inertia control technology on doubly-fed induction generator. In: Chinese automation congress (CAC). IEEE, pp 5503–5507
16. Basha CH, Rani C (2020) Different conventional and soft computing MPPT techniques for solar PV systems with high step-up boost converters: a comprehensive analysis. Energies 13:371
17. Batzelis EI, Anagnostou G, Cole IR, Betts TR, Pal BC (2019) A state-space dynamic model for photovoltaic systems with full ancillary services support. IEEE Trans Sustain Energy 10(3):1399–1409
18. Basha CH, Rani C, Odofin S (2018) Analysis and comparison of SEPIC, landsman and zeta converters for PV fed induction motor drive applications. In: International conference on computation of power, energy, information and communication (ICCPEIC). IEEE, pp 327–334

Detection of Renal Calculi Using Convolutional Neural Networks

A. Madhavi, M. Harshitha, M. Deepak Sai, and N. Anand

Abstract Renal calculi alias kidney stone is a prevalent aching disease. Approximately 15% of the global population is a victim of this illness. Identifying the presence of the stones in the initial stages is a tedious and challenging task. The contemporary solutions for this problem fail to accurately detect the location and measurements of the stone. This project aims at providing a potential solution to the above problem. A combination of CNN models namely XResNet and FasterRCNN is proposed in this paper. The position and measurements of the stones are identified with high accuracy in the Region of Interest of the CT scan images. This helps doctors to suggest appropriate treatment without further ado.

Keywords Classification · Convolutional neural networks · Accuracy · ResNet

1 Introduction

The renal calculi generally called kidney stones or urinary stones are the solid deposit of minerals in the kidney. Low consumption of water, high sodium intake, etc., are few of the multiple reasons that cause this problem. Despite many medical innovations to avoid this illness, this problem is rife all around the world. Though it is easy to treat small stones, it is difficult to locate them. In the current world, CT scan, MRI scan, Urine test, and Blood test are discrete ways to diagnose kidney stones. However, CT scan is more widely used. The CT technology has been a reliable resource for physicians to identify kidney stones. This project is developed intending to automate the detection of renal stone using image processing techniques on the CT scan images. A novel coalescence of XResNet and FasterRCNN is used for performing the classification and stone localization, respectively. Using these CNN models enhances the feature extraction ability when compared with other segmentation techniques

A. Madhavi (✉) · M. Harshitha · M. Deepak Sai · N. Anand
Department of Computer Science, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India
e-mail: madhavi_a@vnrvjiet.in

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_7

like threshold-based segmentation, etc. FasterRCNN on the other hand is an efficient enhancement over FasterRCNN and has a Region Proposal Network.

2 Literature Survey

2.1 Kidney Stone Detection with CT Images Using Neural Network [1]

The CT scan images are preprocessed using the DWT (Discrete Wavelet Transform) method. Features extracted using GLCM feature extraction methods are used to classify the image by the Back Propagation Network (BPN). Further, this system uses a segmentation method to detect the position of the stone. The Fuzzy C-means algorithm is a clustering technique used here for segmentation. The stone location is then obtained from the segmented images, and further analysis is made.

Limitations: BPN is a regular neural network which is sensitive to the noisy data. Furthermore, it takes time to train if the data is large. DWT, on the other hand, is a complex preprocessing technique which is sensitive to minor adjustments of its parameters.

2.2 Renal Stone Detection and Analysis by Contour-Based Algorithm [2]

This existing model requires ultrasound images to perform renal stone detection. This system uses GAC segmentation for stone localization and size detection. Required preprocessing including adaptive equalization and morphological operations is done. Features are extracted from the segmented images, and then, contours are generated in two stages. Then, the regions from the segmented contours are classified as stone or not.

Limitations: This process uses ultrasound images which in general consist of more speckle noises than CT images. Contours only produce the outlines of the stones. There is a chance of being sensitive to detect two or more stones together as one.

2.3 Accurate Kidney Segmentation in CT Scans Using Deep Transfer Learning [3]

The model proposed in this paper employs a U-net which is obtained from the open-source framework. The model is then trained against the dataset containing CT images acquired over 3 years. Fivefold cross validation is employed with over 500 epochs.

Limitations: The method is really sensitive to minor changes in the scan images and mistakes the surrounding anatomy for the stone.

Other than the above-known models, various other papers were explored too. Those observations are summarized in Table 1.

Table 1 Previous research investigation on detection of renal calculi

Year and reference	Paper title	Objective	Techniques	Limitations
2019 [4]	Urinary stone detection on CT images using deep convolutional neural networks: evaluation of model performance and generalization	To identify the stones and analyze the model's performance for their accuracy	Gray nets and ImageNets, CNN	Not enough accuracy
2020 [5]	Automatic detection and scoring of kidney stones on noncontrast CT images using S.T.O.N.E. nephrolithotomy: combined deep learning and thresholding methods	A deep learning system that detects kidney stones is developed and also validated	Cascaded 3d U-net, thresholding	Not enough preprocessing done before feature extraction
2019 [6]	Kidney disease detection and segmentation using artificial neural network and multi-kernel k-means clustering for ultrasound images	Neural networks and segmentation methods are used to detect the kidney stones	Multi-kernel k-means algorithm, Median filter	ANN's suffer from dimensionality curse on large data
2018 [7]	Computer-aided detection of ureteral stones in thin slice computed tomography volumes using convolutional neural networks	Use of CNNs to detect Urinary stones	CNN	Too many false positives, no proper preprocessing is done

(continued)

Table 1 (continued)

Year and reference	Paper title	Objective	Techniques	Limitations
2018 [8]	An image preprocessing method for kidney stone segmentation in CT scan images	Preprocessing algorithms for kidney stone detection and segmentation in CT images were analyzed and the soft, hard organs and blood vessels were removed from the image	Thresholding algorithms	Lacks robustness
2018 [9]	Model for predicting the risk of kidney stone using data mining techniques	Detection of kidney stones considering genetic data among Nigerians	C4.5 decision trees algorithm, genetic programming, and multi-layer perceptron	Focusing entirely on a race and data collection is difficult considering the number of parameters
2019 [10]	Analysis and identification of renal calculi in computed tomography images	Kidney stone detection using digital image processing	Pre-processing, morphological operation, binarization	Limited dataset and not enough accuracy

3 Methods

3.1 Convolutional Neural Networks

Tasks like object detection that require complex structured arrays and vectors have made good use of Convolutional Neural Networks. Other Computer Vision such as Image Recognition also uses Convnets to the maximum extent as they produce patterns (features) from the images. The combination/sequence of various layers such as Convolutional Layer, Pooling Layer, and Fully Connected Layer each performing a distinct action yet work toward detailing key features from an image makes the performance of CNNs more accurate. Each Conv Layer internally is a collection of "Neurons"—the unit components that learn and predict outcomes. A Pooling Layer acts on outcomes of neurons and performs either Max Pooling or Average Pooling usually. A typical Artificial Neural Networks (ANN) comprise 3 layers, i.e., input layer followed by hidden and the output layers. A connection between the hidden units and all the input units and one between the output and hidden units makes it a fully connected network. For instance, an "input unit" (Ij) with a "bias" (bj), and a "weight matrix" (Wij), the "hidden unit" (hi) is calculated as in Eq. (1).

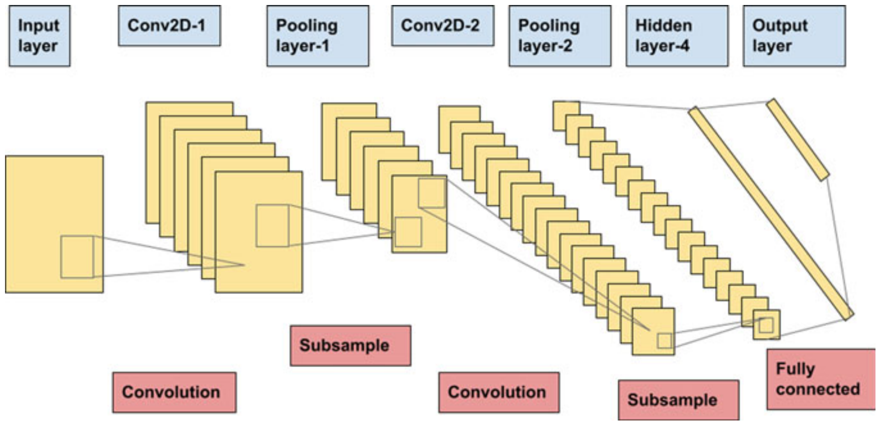


Fig. 1 Working of a ConvNet

$$\text{hi} = \sum\nolimits_{j=1}^{J} WijIj + bi \tag{1}$$

However, when high dimensional inputs come into picture, the model suffers from dimensionality issues. Hence, ANNs are not preferred when we consider object detection.

CNNs, on the other hand, require fewer parameters as it uses local connections with tied weights [11]. It also takes advantage of the fact that input data has a spatial structure. It can also learn from slightly translational and rotational invariant features. A ConvNet can successfully capture both spatial and temporal dependencies in an image and hence is preferred over an ANN.

Figure 1 shows the overview of how image features are extracted and then classified using a ConvNet. It involves a series of Conv2D and Pooling Layers which perform Convolutions and sub-sampling on the output from the previous layer. The size of the image reduces and the number of channels increases as more and more layers are introduced. A Fully Connected Layer and the Classifier are attached at the end usually for the classification purpose.

Classification of CT scans here is done with the use of XResNet-50 which is an improvement over the most commonly used ResNets, a type of ConvNets majorly used in various applications of deep learning.

3.2 *ResNet and XResNet*

Many CNN-based architectures came into existence in the recent years and one that caught everyone's eye is AlexNet which won the ImageNet competition almost a decade ago. This resulted in everyone including more and more layers in their deep neural networks to reduce the error rate. However, the result was not as expected

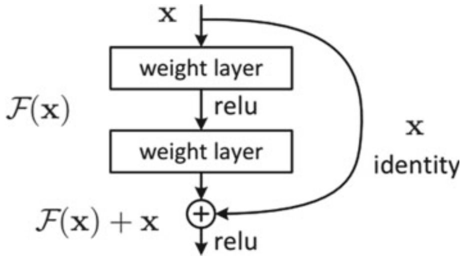


Fig. 2 Skip connection depiction

as the vanishing/exploding gradient problem caused networks with a high number of layers to fail. The concept of residual networks was then introduced to address this problem. It makes use of skip connection technique, i.e., when a layer hurts the performance of the network; it directly skips the training from those layers by regularization and directly connects itself to the output. Figure 3 depicts the ResNet architecture in general. The output image size after a convolution is determined as per Eq. (2)

$$\text{Output size} = [n + 2p - f]/s \ + 1 \tag{2}$$

where n is input size of the image, f is the kernel size, p is the padding applied and s is the stride (Fig. 2).

There are many forms of ResNet based on the number of convolution layers and other factors including ResNet-18, ResNet-34, ResNet-50, and others. ResNet-50 contains 48 convolutional layers, 1 max pooling layer, and 1 average pooling layer.

XResNets are introduced to squeeze a little more performance from your ResNet. Minor tweaks [12] are done to the original ResNet architecture in order to improve the efficiency. This algorithm observes adjustments to the regular ResNet architecture to improve the learning rate decay, mix-up training, label smoothing, and knowledge distillation. ResNet B, ResNet C, and ResNet D are the modifications on the ResNet architecture which basically are changes of particular convolutional layers and changes in strides. This makes the XResNet model less computationally expensive and so training this model is easy (Fig. 4).

3.3 FasterRCNN

Localization of a stone after classifying the image is where FasterRCNNs come into play. FasterRCNNs can have various feature generation algorithms as its base including vgg-16, ResNet, etc. ResNet-50 is a base algorithm which generates the feature maps from the input image. These are then passed to the Region Proposal Network which as the name suggests generates various region proposals with the use of another ConvNet. As you can see in Fig. 5, the RPN is connected to the

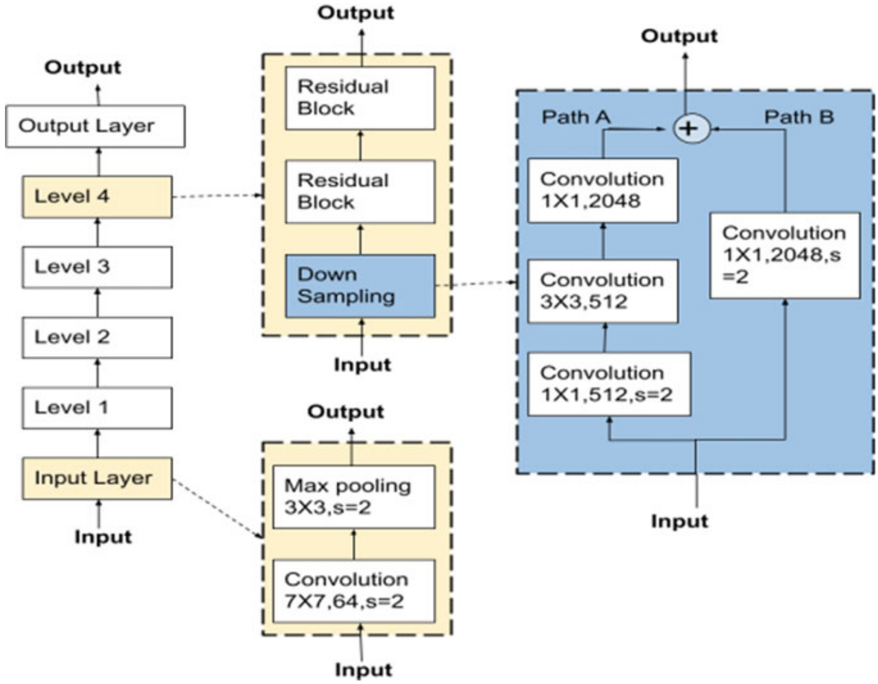


Fig. 3 ResNet architecture

Convolutional layers (3 × 3 followed by 1 × 1 to know whether the region has any object or not) and in turn to the softmax and regressor. ROI pooling is next applied to the images to the proposed regions, and then, the output is connected to a few fully connected layers to flatten them. Finally, they are subjected to a softmax function and then to linear regression for classification and to draw the boxes, respectively. The model outputs the coordinates along with the boxes around the stone(s) on the image and approximate length, width, and area of the stone(s).

3.4 Proposed Methodology

The process proposed in this paper is represented in the form of a flowchart in Fig. 6. The preliminary step involved is preprocessing the input CT scan image.

In this stage, the main focus is to reduce the speckle noise that is usually present with the CT scan images. Speckle is an undesired effect that comes with the scan images. They disturb the visual and digital information of the image and make it complex to understand. In order to perform the job of removing the noise from the image, Bilateral Filter [13] is used here. This filter smoothes the image and also retains the edges. While the regular filters like Gaussian filter and median filters

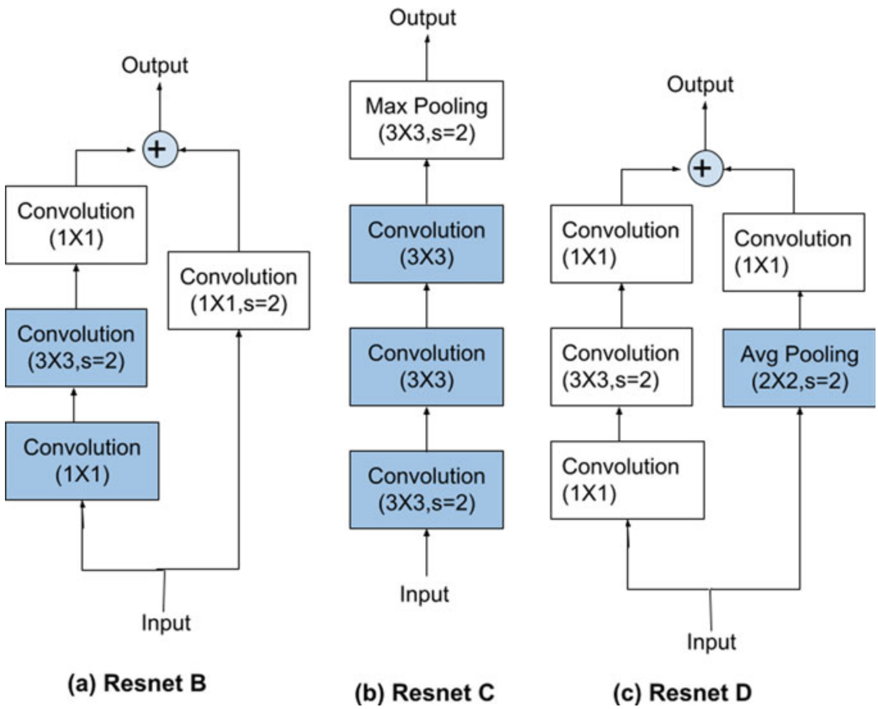


Fig. 4 Tweaks of ResNet

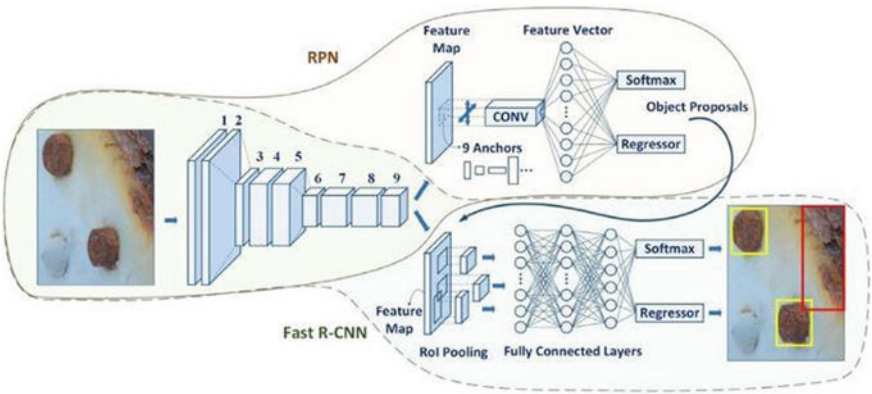


Fig. 5 FasterRCNN

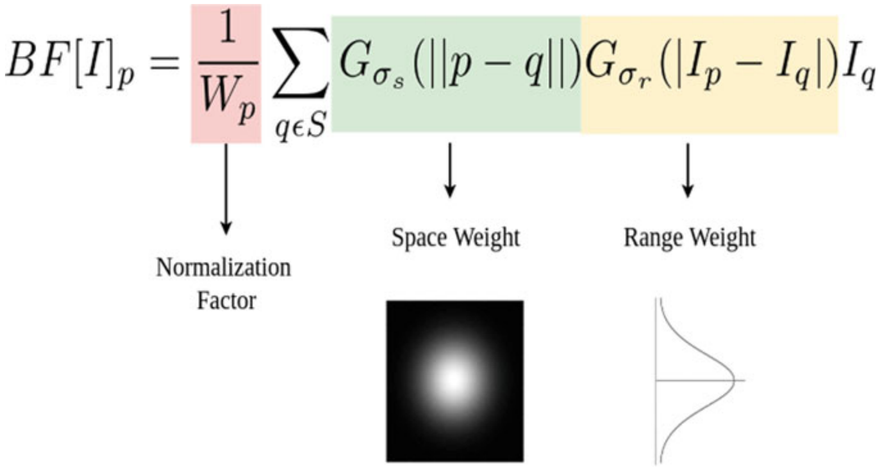


Fig. 6 Process of proposed method

blur out everything irrespective of the edges and objects thereby losing important information, bilateral filter is non-linear and uses normalization factor and also range weight in addition to the space weight. Below is the formula of the bilateral filter as presented in Eq. (3).

Here, σ_s is used to denote the kernel's spatial extent, i.e., the neighborhood size to be considered. Smaller its value indicates sharper edges. σ_r is used as the minimum amplitude of the edge. It seems that only pixels with similar intensity as that of the central pixel get blurred and avoid sharp intensity changes. Also, $G_{\sigma s}(\|p - q\|)$ is the normalized Gaussian function.

Now that the preprocessing is done, we go ahead and feed the images to the XResNet model. We train the model with CT scans and then once the model is trained, we go ahead and test the model for its accuracy. If an image is fed by the user to model and if the image classifies the image as the one with the kidney stone, we go ahead for the further step which is localizing and determining the size of stone.

The training of FasterRCNN is a little different to that of ResNet. In ResNet, we label the images whereas for FasterRCNN we annotate the image with stone locations which are supplied to the model in the form of a json file. After training these models, testing and validating of the model follow. Now that we have the bounding box coordinates of stone in image, size can be found out contouring the exact shape of the stone within the dimensions.

Table 2 Dataset distribution of CT scans

	Train	Test	
Stone	625	165	790
Normal	828	181	009
	1453	346	1799

4 Experimental Results

4.1 Experimental Data

The dataset consists of 1799 CT scan images of which 790 images contain one or more kidney stones. The study group included scans of patients of varied types including those lacking a kidney or with urethral stones. The weights of the included are not known and neither are their details. Dose modulation has not been done, and hence, the noise is the result of the body size. A solitary or a conglomerate of stones is present in 790 scans. The positions of the stones present are annotated for all the scans and are approved by a radiologist. The annotations file saved as the json file serves as an input to the network along with the data. The dataset structure as shown in Table 2 encompasses train and test folders which further are labeled as "stone" and "normal." The size of each CT scan is varied and is huge in regard to processing as abdominopelvic scans are considered instead of the regular ultrasound images.

4.2 Classification

The model is evaluated on the training set and later on the validation dataset for every epoch. Learning curves on performance are drawn to evaluate the occurrence of any issues such as underfitting and overfitting while training. In the graph depicted, no such problems are encountered as train loss is always less than or equal to validation loss (Fig. 7).

Once the model is tested against the test dataset, confusion matrix is plotted to determine the false-positive rate, false-negative rate, and accuracy as per the following Eqs. (3), (4), and (5) (Fig. 8).

$$\mathrm{FPR} = \mathrm{FP}/\mathrm{N} = \mathrm{FP}/(\mathrm{FP}+\mathrm{TN}) = 1 - \mathrm{TNR} \tag{3}$$

$$\mathrm{FNR} = \mathrm{FN}/\mathrm{P} = \mathrm{FN}/(\mathrm{FN}+\mathrm{TP}) = 1 - \mathrm{TPR} \tag{4}$$

$$\mathrm{ACC} = (\mathrm{TP}+\mathrm{TN})\,/\,(\mathrm{P}+\mathrm{N}) = (\mathrm{TP}+\mathrm{TN})/(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}) \tag{5}$$

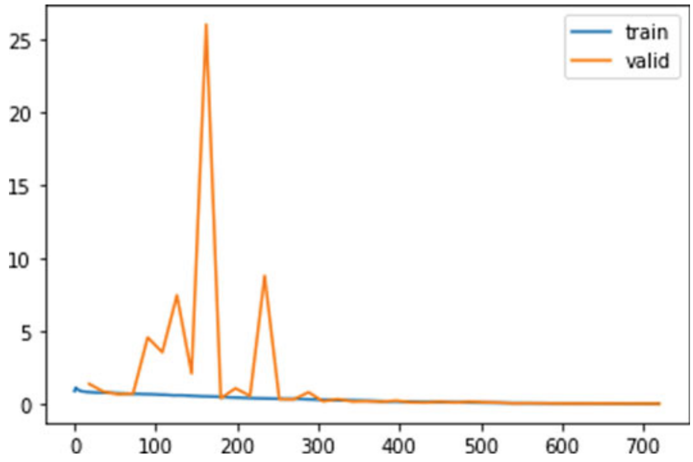
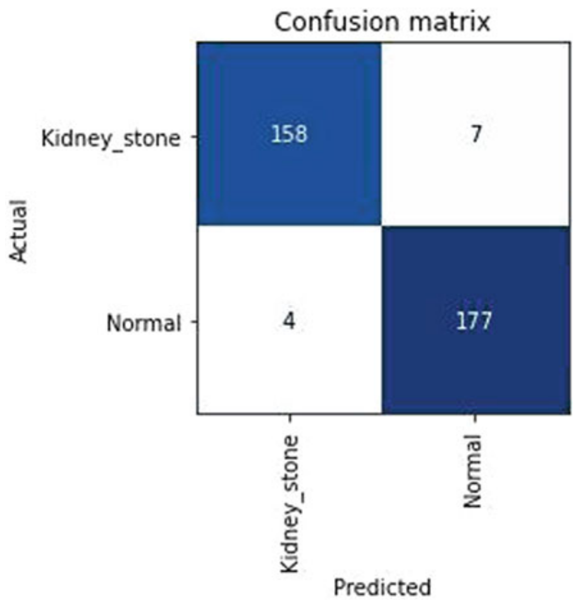


Fig. 7 Plot depicting train and validation loss

Fig. 8 Confusion matrix on test set



where FPR is false-positive rate, FP is false positives, N is total negative, TN is true negative, TNR is true-negative rate, FNR is false-negative rate, FN is false negatives, P is total positive, TP is true positive, and TPR is true-positive rate.

From the above calculations, FPR is 0.0221, FNR is 0.0424 and accuracy is 96.82%.

Table 3 depicts the classification report of the trained model.

Table 3 Classification report

	Precision	Recall	F1-score	Support
Kidney_stone	0.98	0.96	0.97	165
Normal	0.96	0.98	0.97	181
Accuracy			0.97	346
Macro avg	0.97	0.97	0.97	346
Weighted avg	0.97	0.97	0.97	346

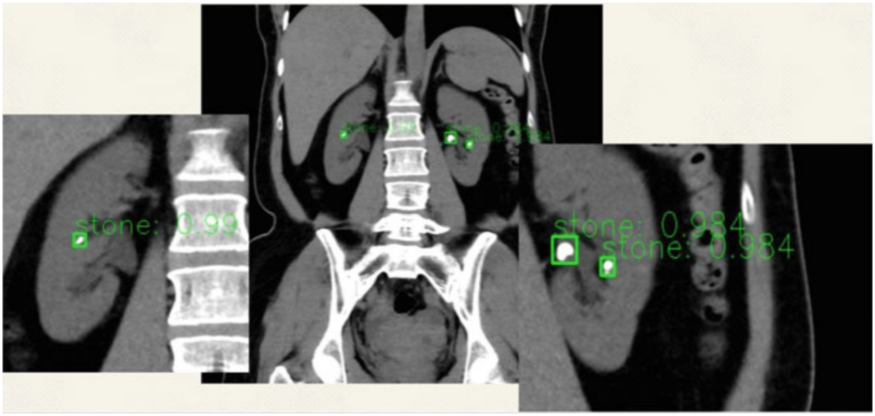


Fig. 9 Resultant image after localization

4.3 Localization

Figure 9 shows the prediction of the model on a test image. The outcome of the model depicts the bounding box over all the stones present in the CT scan and a score associated with it which is the confidence score. The confidence score for a bounding box tells us with how much confidence the model is saying that the pixel in the bounding box predicted is a stone. The confidence limit is set to 75%. All the predictions below it are not considered.

5 Discussion

All the images consisting of kidney stones from both test and train datasets are manually annotated using VGG annotator to train and test the localization algorithm. An approval of those positions is taken from a radiologist. FasterRCNN is implemented in PyTorch on a GPU provided by Google Colab.

6 Conclusion and Future Work

In the future, the model can be extended to classify and detect stones based on their position such as urinary stones present in the urinary tract and renal stones located inside the kidney region. Furthermore, the work can be made easily accessible with a user interface.

In this paper, we sought to determine the size and location of the kidney stone using CT scan images. A novel combination of algorithms is used. The classification and localization done by XResNet-50 and FasterRCNN gave good results with high accuracy. Also, the model performed well on unseen test subjects. In conclusion, the model can be used to accurately predict the position and size of the renal calculi using CT images. This helps the doctors to assess the severity of the condition and treat the patient accordingly.

References

1. Mishr R, Bhattacharjee A, Gayathri M, Malathy C (2020) Kidney stone detection with CT images using neural network. Int J Psychosocial Rehabil 24(08)
2. Shahina MK, Mahesh HS Renal stone detection and analysis by contour based algorithm. 978-1-5386-9482-4/19/$31.00 ©2019 IEEE
3. Graham-Knight JB, Scotland K, Wong VKF, Djavadifar A, Lange D, Chew1 B, Lasserre P, Najjaran H (2020) Accurate kidney segmentation in CT scans using deep transfer learning. In: McDaniel T et al (eds) ICSM 2019, LNCS 12015. Springer Nature Switzerland AG, pp 147–157
4. Anushri Parakh MD, Hyunkwang Lee MS, Jeong Hyun Lee MD, Eisner BH, Dushyant MD, Sahani V, Synho Do MD (2019) Urinary stone detection on CT images using deep convolutional neural networks: evaluation of model performance and generalization. Radiol Artif Intel 1(4)
5. Cui Y, Sun Z, Ma S, Liu W, Wang X, Zhang X, Wang X (2021) Automatic detection and scoring of kidneystones on noncontrast CT images using S.T.O.N.E. nephrolithotomy: combined deep learning and thresholding methods. Mol Imaging Biol 23:436–445. Published 27 Oct 2020
6. Nithya A, Appathurai A, Venkatadri N, Ramji DR, Anna Palagan C Kidney disease detection and segmentation using artificial neural network and multi-kernel k-means clustering for ultrasound images. Elsevier/locate/measurement, Issue 19/08/2019
7. L¨angkvist M, Jendeberg J, Thunberg P, Loutfi A, Lid'en M (2018) Computer aided detection of ureteral stones in thin slice computed tomography volumes using convolutional neural networks. Comput Biol Med 97:153–160
8. Thein N, Nugroho HA, Adji TB, Hamamoto K An image preprocessing method for kidney stone segmentation in CT scan images. In: 2018 international conference of computer engineering, network and intelligent multimedia (CENIM)
9. Oladeji FA, Idowu PA, Egejuru NC, Faluyi S (2019) Model for predicting the risk of kidney stone using data mining techniques. Int J Comput Appl (0975–8887) 182(38)
10. Shah T, Kadge S Analysis and identification of renal calculi in computed tomography images. In: 2019 international conference on nascent technologies in engineering (ICNTE 2019) 978-1-5386-9166-3/19
11. Yamashita R, Nishio1 M, Do RKG, Togashi K (2018) Convolutional neural networks: an overview and application in radiology. Insights into Imag 9:611–629

12. He T, Zhang Z, Zhang H, Zhang Z, Xie J, Li M Bag of tricks for image classification with convolutional neural networks, Amazon Web Services
13. GeeksforGeeks, Python|BilateralFiltering, published Sep 2019 https://www.geeksforgeeks.org/python-bilateral-filtering/

Question Answering and Text Generation Using BERT and GPT-2 Model

Santoshi Kumari and T. P. Pushphavati

Abstract Question and answering system and text generation using the BERT and GPT-2 transformer is a specialized field of the information retrieval system, which a query is stated to system and relocates the correct or closet answer to a specific query asked by the natural language. The main aim of the QA system is to provide the short answer to a question rather than the list of possible relevant document and text generation is a language generation which focuses on the producing understandable test in English which can predict the next sentence or generate a text with all the raw content from previous words. The motivation for selecting this work is to provide a great relevance to find the answer, find answer to general knowledge type of question, find the answers for questions like Who? What? Where? How?, and Provide provide the shortest form of answer. The scope for the chosen work is to provide the solutions for the automation moderation in the websites to provide the exact and short information answers from the websites, like Stack Overflow, Reddit, Quora, provide the self-answering and find text. The method we are using for the QA and text generation system is a transformer architecture which consist of Encoder and Decoder which is a stack of encoder represents the BERT model and Decoder part is represented as the GPT-2 model.

Keywords Text generation · BERT model · GPT-2 model · Question and answering · Transformer · Question · Natural language processing · Encoder · Decoder

Please note that the LNCS Editorial assumes that all authors have used the western naming convention, with given names preceding surnames. This determines the structure of the names in the running heads and the author index.

S. Kumari (✉) · T. P. Pushphavati
Department of Computer Science and Engineering, M S Ramaiah University of Applied Sciences, Bangalore, India
e-mail: santoshik29@gmail.com

T. P. Pushphavati
e-mail: acepushpa@gmail.com

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_8

1 Introduction

The transformer models have becoming more powerful in field of natural language processing and transformer architecture is enabling the large scale in language model to train on large amount of data to improve the state of the art in NLP task and with these models, we are leading the significant amount of improvements in the language processing tasks and with hundred millions of parameters and research papers on the pre-trained models indicates the better training and improved in the performance of the models and created a number of benchmarks. In the past several years, question answering websites have been very useful for many users and have been a reliable source of information in the various fields. The main goal of our project is to providing interface for the users with the exchange and share of knowledge from the questions arrived [1]. If the user asks a question that lacks the knowledge on a specific topic or cannot find the desired answer to the question our model predicts the answer related to the question. In this way, the QA will be a source of information that will replace the other source like database or documents. The model which we are using is Bidirectional Encoder Representation Transformer (BERT) [2] which is unsupervised learning approach. The technical feature of BERT model is to apply the bidirectional training to the transformer to a language model and this provides the text sequence to look from both left to right and right to left which consists of the two separate mechanism which is an Encoder reads text input and Decoder which produces the prediction result. For the observation in training data sets, we have a question context and a text and our goal is to find answer from the context provided and we are using SQUAD 1.1 data sets [3] which contains 40 paragraph questions with 100,000+ question and answer in pairs on the 500+ articles [3]. The question answering provides the community for websites like Quora, Stack Overflow with their relatable answer in the comment section and provides the automated modern actions to the QA websites.

Language modelling has become important field in the advance natural language processing in which the language model is the probabilistic model which is used for the predicting the next word or a character in a text format or generate a text, so one application in the large-scale neural network model has emerged the NLP approach which is text generator with the pertaining the massive amount of the unlabelled text, and these model has capable of achieving the state-of-the-art performance in the NLP benchmarks. The one particular model is the GPT-2 model is developed by the OpenAI [4]. The full GPT-2 model consists of 1.5B parameters and model trained from the data of 8 million web pages collected from the links of Reddit [4]

The text generator is used to predict next word or sentence with all the given previous words within some context. The variety of the data set had made a simple goal to contain the natural occurring demonstration with various task and domains, and GPT-2 has been scaled with 10× parameters trained on 10× amount of data [4].

2 Related Work

Linguistic Approach: The question answering system has to understand natural text, linguistic knowledge, linguistic techniques such as parsing, tokenization and POS tagging were implemented to the users question for the precise query to extract the structured data. The specific domain knowledge requires a different mapping rules and grammar to build a best system and to build that it requires a time-consuming process.

Statistical Approach: It is a structured query language which can formulate the queries of natural language forms, and it needs a huge amount of data for statistical learning once it needs to learn and produces the better results. The statistical approaches have been applied in different stages of the QA system such as support vector machines (SVM), maximum entropy models, and Bayesian classifiers. These are the techniques used to analyze the question and make the prediction about the correct and suitable answers. The system uses the maximum entropy model for the question answering system based on the bag-of-words features or N-gram, and SVM is used the SVM classifier based on the features of words, named entity, POS tagging and implements the SVM classifier [5].

Word Embedding (Word2Vec): The word2vec model using the skip gram is advance than the bag-of-words and it contains two ways to do skip gram and continuous bag-of-words (CBOW). [6] Skip gram model adds the word vectors and gives the meaningful phrases or sentence and it is pre-trained the large volume of text and it identifies the text and also vector represents for the word of similar meaning for example 'need' and 'require' will have cosine similarity and word2vec understands much better than bag-of-words.

Glove Embedding (Global vector for word Representation): It has been developed similar to the word2vec but it is trained differently and it uses the matrix factorization a technique of deep neural network and creates word-to-word coherence matrix uses the factorization methods to understand the word context [7]. In this paper, the model produces the vector space with substructure of meaningful value, and it performed the 75% on the word analogy and outperformed the named entity recognition [7].

Long Short Term Memory (LSTM): It is an artificial recurrent neural network which consists of feedback connection and can be processed an individual data points and entire sequence of data. In question answering, here it is used the SQuAD v1.1 [3] data sets.

The method [8] used is Match-LSTM model used to predict the text entailment where two sentences are given in which one is premise and other is hypothesis. To predict the entails and hypothesis, the LSTM goes through the tokens sequentially. At each point of hypothesis, it goes through weighted vector of premise. The weighted vector combined to vector representation to current token and fed to LSTM which we can match the sequence of token data. Here they first tokenize all question and answers passages [8]. The vocabulary contains the 117 k unique words and uses the word embedding GLOVE to initialize the model.

Text Generation Using Generative Adversarial Network (GAN):

SeqGAN: The techniques they have used in the SeqGAN for generating the Barak Obama speeches and Chinese poems [9]. They have used the corpus Chinese quatrains and containing the twenty characters of the Chinese poems [9]. A corpus of 11,092 paragraphs were used from past Obama's political speech for generation of new political speech.

3 Methodology

Question answering framework consists of three modules which are:

1. Query processing module
2. Document processing module (information retrieval)
3. Answer processing module.

The main goal of the system is to extract the number of pieces of information from question. Fig. 1 [10] illustrates the question with their processing using where the question is passed into the information retrieval and query specifies the keyword which is used for the information retrieval which is used for searching for the document or paragraph [11].

- Analysing the query for the information which is required for the answering query.
- Classifying the questions with their keywords used and taxonomy, which helps in identifying the answer type.

3.1 Model Architecture

Transformer encoder uses the attention mechanism which learns the relationship between the words and sub words in text. Transformer uses encoder-decoder architecture. Fig. 2. Illustrates the Encoder Decoder transformer layer where the encoder

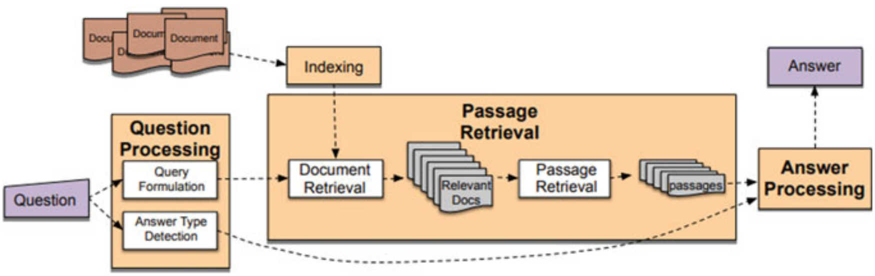


Fig. 1 Question answering system framework

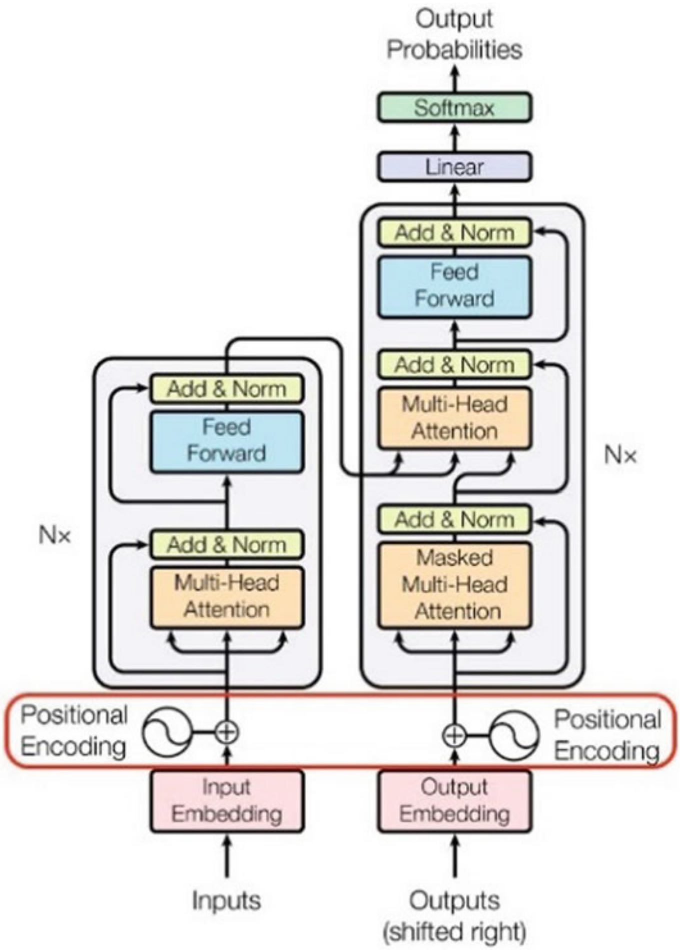


Fig. 2 Transformer architecture

layers contain of stack of six encoder. Encoder layer consists of two layers: self-attention and feedforward neural network. The attention mechanism helps current nodes to focus on current word and obtain the semantic context of information. The decoder contains two-layer network same as encoder and contains the layer in middle which is attention layer and help the current node for getting the key information that needs for attention as show in Fig. 2 [12].

The Encoder [12] is on the left and the Decoder is on right. Both of them are composed of models which are stacked on top with multiple times described by Nx in figure. As we observe the transformer model mainly consists of feedforward layer and multi-head attention layer, the target sentences such as input/output are embedded into n-dimension space. These Encoder is represented as the BERT model and Decoder layer is represented as the GPT-2 model derived from the transformers.

3.1.1 Question Answering Using BERT Model

Bidirectional Encoder Representation Transformer (BERT) which is developed by the Google and published in 2018 by Jacob Devlin and his colleagues from the Google [2]. BERT uses the transformer encode block which transformer uses the Multi-Headed Self Attention mechanism which learns the contextual information within the words in the text. The masked language model is pre-trained objective and randomly mask the tokens from input and predict the objective of masked word and its surrounding. The sequence of the model is bidirectional which is right to left and left to right and allows the pre-training direction [2].

The BERT model reads input text form left to right and right to left sequentially and transformer reads the entire sequence at once helps in better accuracy than non-directional as shown in Fig. 2.

3.1.2 BERT Architecture for Question Answering System

The BERT produces the input query passage pair and additional layer on top of it [3]. As shown in Fig. 3, the model is designed for the SQuAD Data set of question answering, and the layers are created for the contextual representation of the input query passage pair; the model will start with the pre-training of the BERT weights and fine-tuning of the training data.

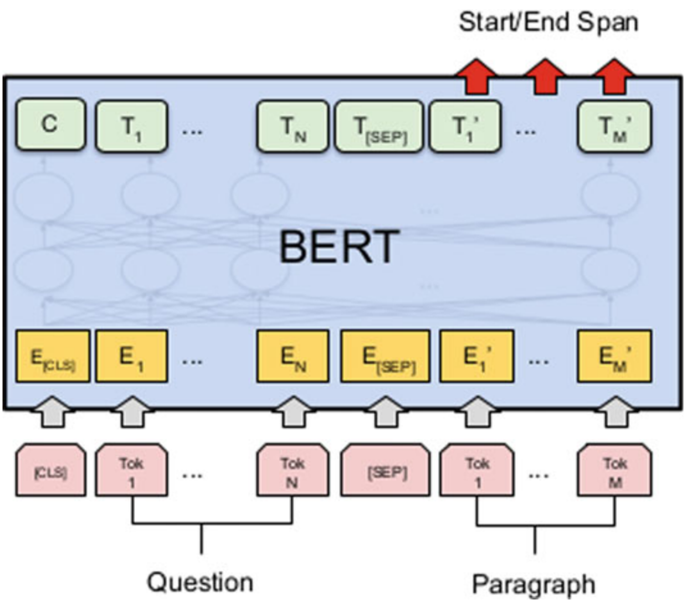


Fig. 3 BERT question answer

BERT Architecture:

- BERT small (Base): Consists of 12 layers (transformer blocks), 110 million parameters and 12 attention head
- BERT Large: consist of 24 layers, 340 million parameter and 16 attention head.

As for illustration, Fig. 4 represents model BiDAF contains five layers with high-level structure which can be found in SQuAD data sets [13].

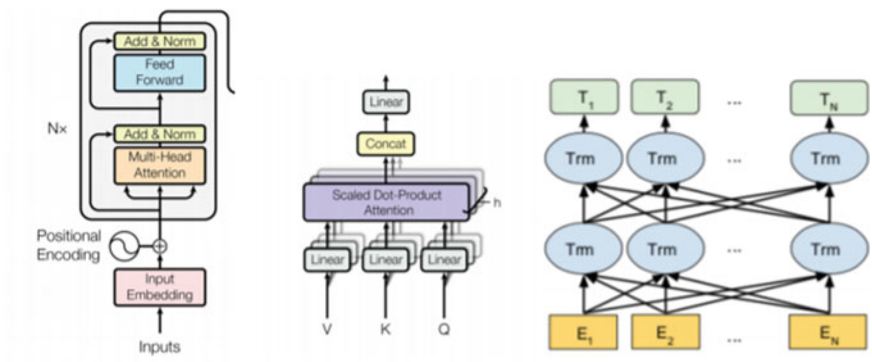
1. Embedding Layer: It converts the indices into the GloVe word embedding vector with embedding size of 300 [13].
2. Attention Layer: BiDaF layer flows the information from the question context.
3. Encoder Layer: Two layer long short term memory for encoding of embedded sequence.
4. Output layer: It produces the probability vector from start to end position and calculates the SoftMax for output from start to end pointer for the answer span.
5. Model Encoder Layer: It contains of one layer with bidirectional LSTM for refining.

Sequence of vectors after attention layer [13].

The BERT transformer consists of two sub layers, feedforward and multi-head attention mechanism [13].

1. **Multi-Head Attention**: As shown in Fig. 4 the multi-head attention function is described by the query (Q) as a mapping function and the key value pairs of the set (K, V) and output layer (O) and it is a sum of the values. Weights are assigned for each of value and send it to the query (Q_i) with their key value (k_i). And BERT matrix of input X which is an input or hidden layers with their following weight matrixes [13]:

$$XW_i^Q = Q_i \in \mathrm{R}^{\mathrm{input}X\mathrm{d}q};$$



(a) Transformer Encoder (b) Multi-Head Attention (c) BERT Architecture

Fig. 4 **a** Transformer encoder, **b** Multi-head attention and **c** BERT architecture

$$XW_i^K = K_i \in \mathrm{R}^{\mathrm{input}X\mathrm{d}k};$$

$$XW_i^V = V_i \in \mathrm{R}^{\mathrm{input}X\mathrm{d}v}$$

Multi-head attention mechanisms linearly project queries, values h and keys. The linear projection dv, dk and dq dimensions and the multi-head perform the parallel in function and output vector is projected once and resulting the output vector (O) [13].

$$\mathrm{MultiHead(Q, K, V)} = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h W^O);$$

$$\mathrm{head}_i = \mathrm{Atten}(Q_i K_i V_i)$$

2. **FeedForward**: These networks consist of the two linear transformation with the ReLU activation function [13].

$$\mathrm{FFN(O)} = \mathrm{RELU}(\mathrm{O}W_1 + b_1)W_2 + b_2$$

Text Generation using GPT-2:

Language Modelling: It is the field of artificial intelligence and computational linguistics which is an important task of the modern NLP Task [4]. The system focuses on the computer system than can produce the understandable texts in English and generate the text or predict the next sentence.

GPT-2 is the newer version of the GPT which is an NLP framework developed by OpenAI [4] which is used for the prediction of the next work from the 40 GB of the Text from Internet.

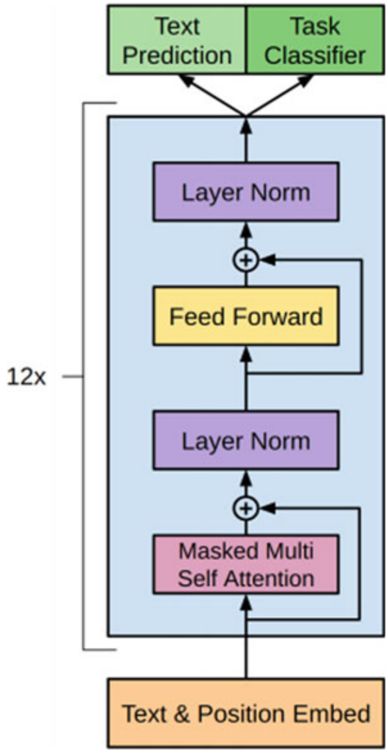
GPT-2 is a Generative Pre-trained Transformer which is a transformer-based model which consists of 1.5 billion parameters and trained on the data sets of 8 million web pages [4], and objective is to generate the text or predict next sentence with all the previous sentences of some context or raw human word and GPT-2 is trained more than 10× amount of data and contains 10× parameters. The pre-trained model contains of 8 million data of web pages which is collected from the links of Reddit [4].

Generative: It means model is trained for the generation or prediction of the next word or token in a sequence format of the tokens in unsupervised learning, it means the model is thrown whole lot of the raw text and it will generate the statistical feature and predict next sentence or create more related text value.

Pre-trained: It is very powerful and large model which is fined-tuned in a specific machine translation.

Transformer: It means it uses the transformer architecture to build a model.

Fig. 5 GPT-2 architecture



GPT-2 Architecture:

As for illustration, Fig. 5 [14] represents the GPT-2 which is an architecture based on the transformer of a Decoder block which has modifications with some parameters and transformer layers:

- Model uses large vocabulary size and context of information.
- After self-attention block, another layer normalization is added
- As building block layer, normalization is moved to the input of every sub-blocks, and then normalization is applied before the weight layer.
- The 'decoder' has multiple layers with same input/output sequence. A single decoder contains a multi-headed masked attention operation and fully connected layer. The full connected layer applied individually to each of the vector in the sequence.

We specify each element in the input/output sequence as vector representation of a word. The terminology for the vector is word embedding and uses the dot product to measure the vector similarities [15] for the specific word embeddings.

Decoder takes the sequence of word embedding as input and produces the sequence of word as output.

Decoder is first trained in unsupervised manner using standard language modelling. The loss which we measure the difference between input word embedding and output embedding as shown in Fig. 5. Decoder is used for the prediction of next word and trained in sequence with all the previous words in the sequence. GPT-2 generates response word by word with individual word sample from final output embedding passing through entire decoder, and input of each pass is sequence of length 1 contains embedding of previous words.

GPT-2 keeps tracks of all past keys and with input vectors and each decoder has one key and input vector from in its length -1 input sequence to other pass and pass produces the single embedding from another word is sampled and thus process repeats, adding single word value to GPT-2 for each pass and predicting the next sentence.

This GPT-2 is a 12$\times$ layer decoder of a transformer with 12 independent attention mechanism which is called 'heads' follows with the result of $12 \times 12 = 144$, as shown in Fig. 5, distinct attention patterns, and each pattern corresponds with linguistic property captured by model.

4 Training Procedure

Pre-processing: The text is tokenized using byte-level version of the byte pair encoding for Unicode characters and vocabulary size of 50,257. Inputs are sequenced of 1024 consecutive tokens.

The smaller model is trained on 256 cloud TPU v3 cores.

The transformer expects a complete sentence which means sequence of sentence of words are having fixed length of 512 tokens case of GPT which encodes transformer using decoder [16].

The machine behaviour is the choice of sequence to sequence application, where we want to generate the next word by their word of sequence in a raw format. The encoding representation of sequence generates the new sequence by word by word.

$$\text{word}_t = \text{Decoder}(\text{word}_{t-1}, \text{encoding})$$

If it throws the encoding part from decoder, we will get this:

$$\text{word}_t = \text{Decoder}(\text{word}_{t-1})$$

Fine-tuning GPT-2:

The prediction of missing words, before training, we should set the bos token and eos token as defined earlier in our data sets. We should also set the pad token because we will be using *LineByLineData set*, which will essentially treat each line in the data set as distinct examples. In *transformers/example/language-modeling/run-language-modelling.py*.

After running this code, the special tokens will be added to the tokenizer and the model will resize its embedding to fit with the modified tokenizer.

We set device per train, batch size = 2 and per device, eval batch size = 2 because of the GPU constraints. Feel free to use a batch size that fits your GPU. We use line by line, which tells our model to treat each line in our data set as an individual example, as explained earlier. Evaluate during training runs evaluation on the evaluation data set after each logging steps, which is defaulted to 500.

5 Results and Discussion

Data set:

- We used Stanford Question Answering Data set (SQuAD 1.0) which is released in 2016 by Rajpurkar et.al. [3] is a reading comprehension data set used to evaluate models and train the models to get the accurate answers. The sample consists of data sets in which it includes context paragraph, question and answers [3]. The questions are posed by crowd workers and contains 40 paragraphs questions. The paragraphs are from Wikipedia articles which contains questions from the paragraph and related answers to it.
- Text Generation use the WebText data sets scraper used to train GPT-2 [4]. The current result is just over 23 million URLs and over 10 million HTML pages. Extracting all the Reddit post URLs from the Reddit submission data sets and links will be reduplicated, filtered for exclude non-html content and randomly shuffled and web pages are extracted using the python package. Using FastText, non-English web pages filtered out and contains of 40 GB of text data from 8,013,769 documents.

Text Pre-processing or Input Representation:

As for illustration, Fig. 6 represents the BERT as question as input as only one sequence of packets [13] and input embedding is sum of all the token embeddings.

1. **Token embeddings**: [CLS] Token is marked as input word token and in start of the question, [SEP] token is inserted end of the both question and paragraph.
2. **Segment embeddings**: This indicates the sentence A and B added to the each of tokens, which allows to distinguish between the sentences. All tokens marked as A belong to the question, and those marked as B belong to the paragraph or reference, as shown in Fig. 8.
3. **Positional embeddings**: This is added to every token to get the positions of the sentences [13].

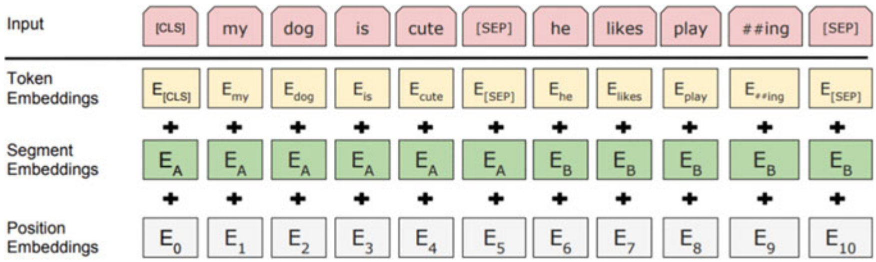


Fig. 6 Text pre-processing of BERT

Pre-training Tasks:

Phase 1

As for illustration, Fig. 7 indicates the pre-training of masked language model is a pre-training task which is the use of taking advantage of masked language model [2] BERT uses the bidirectional as the pre-training task by some of the tokens randomly and using the other tokens to predict the masked token and learns the representation unlike the other approaches, which reads the entire input and predict the masked token.

Before feeding the word sequences into BERT, the experiment picks the 15% of the token randomly to replace and fine-tuning stage and predicts the actual answer.

A: The [MASK] token will be replaced at 80% of the time
B: At 10% of the time, it is replaced by another actual token
C: At 10% of the time, it is kept as original.

Phase 2

Next sentence prediction involves in the understanding the relationship between the sentences and adding the auxiliary task for the training a binary classifier to predict the

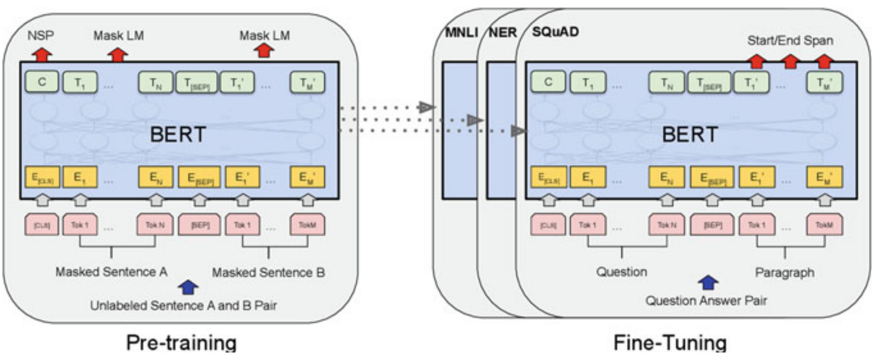


Fig. 7 Pre-training and fine-tuning system

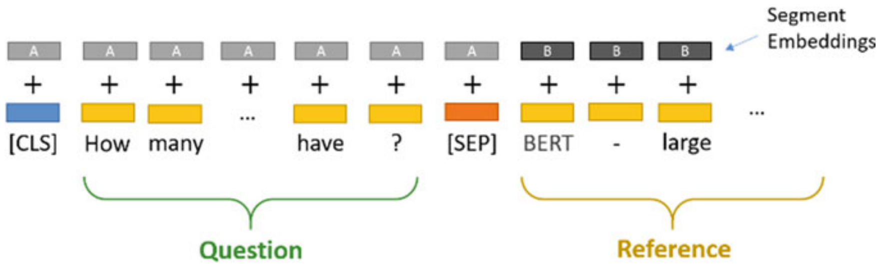


Fig. 8 Fine-tuning system

next sentence to other. During training, 50% of inputs are pairs and second sentence is the original document, and the 50% of the other random sentence from corpus is chosen as second sentence.

Fine-Tuning the Model:

The fine-tuning of the BERT for the question answering system introduces with a start vector and an end vector. Probability of every word will start and calculated by taking the dot product between the final embedding of word and start vector, and the probability of token I is the answer start and which is computed by the SoftMax (S, K), start vector and K is the final output of transformation token I and an end token. The word with the highest probability value is marked as the final answer for the question as shown in Fig. 8.

Add small weight matrix W:

$$h_L^{[i]} w_{\text{SPAN\_START\_END}}$$

Question is impossible to answer if:

$$\text{Max}(\text{start}) + \max(\text{end}) - (\text{Start}_{[CLS]} + \text{end}_{[CLS]}) < \text{impossible\_threshold}$$

Experimental Result for Question Answering System:

- The input is the question which is an article about the Super_Bowl_30 which we have taken and converted all the data into the. JSON format into the rows and columns and pre-processed the functions using the lemmatization, stop words and tokenization for the better understand the answer
- To obtain the question and answer, we have text pre-processed the data by the [CLS] token for the question and [SEP] token for the paragraph for the input words in the token of the question.
- The fine-tuned BERT process the start vector and an end vector. For each vector, the start-word is calculated by dot product between the final embedding of start vector, then SoftMax is calculated with the highest probability values are considered.

- Training Phase: Google Collaboratory uses either Nvidia T4 GPU or Nvidia K80 GPU. The T4 is slightly faster than the old K80 for training GPT-2 and has more memory allowing you to train the larger GPT-2 and BERT models and generates results Nvidia 940MX from my personal laptop.
- To secure our data and minimize time consumption, we took necessary steps.
- Hence, we tested for every 200 iterations to display the number of iterations trained and for every 1000 iterations a checkpoint, Meta file and index file were generated to store the temporary data in Google drive for further usage.
- Implemented all these in form of Android application as GUI
- We trained our model on the data set of SQuAD 1.1 contains question answer and 500 articles for the testing and training. We have split 80:20, 80 for the training and 20 for the testing the data sets and 48 question answering paragraph for the validation.
- For every 50,000 iterations, we used the above generated the text and answers for the questions and if it gives the good result and accurate ground truth result, we would continue checking and we went on the 100,000 to get the satisfactory result.

Results for Question Answering System Android Application:

See Fig. 9.

Results for Text Generation Android Application:

See Fig. 10.

Performance Evaluation: To evaluate the performance, we uses the standard SQuAD performance metric:

Exact Match (EM): Binary measures of the system where the output is matched with their ground truth answer.

F1 score: It is a harmonic mean of the precision and recall.

Performance of the model trained on the SQuAD data sets as shown in Table 1 (Fig. 11).

6 Conclusions

In this work successfully developed the question answering system using the SQuAD [3] data sets by applying the BERT model and applying the variety of the hyperparameters and learning rate, training epochs and different activation functions. The DistilBERT which is a pre-trained version of the BERT consists of 60% faster and 40% smaller and it has better understanding in the language capabilities with 97% of the accuracy [17]. The SQuAD data set consists of the articles of the crowd workers and Wikipedia which consists of question and their ground truth answers we are getting almost every answer from the data set, and the model is achieving the best

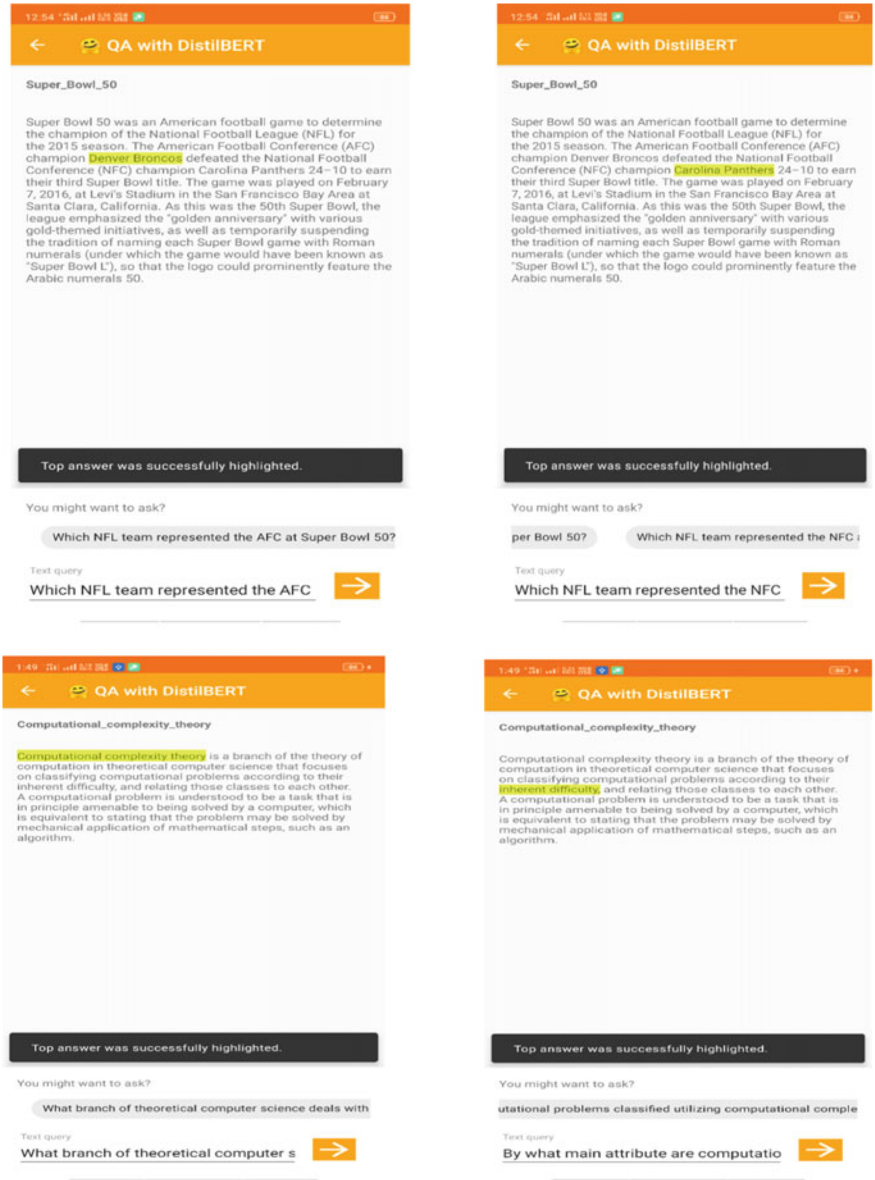


Fig. 9 Highlighted mark shows the predicted answers

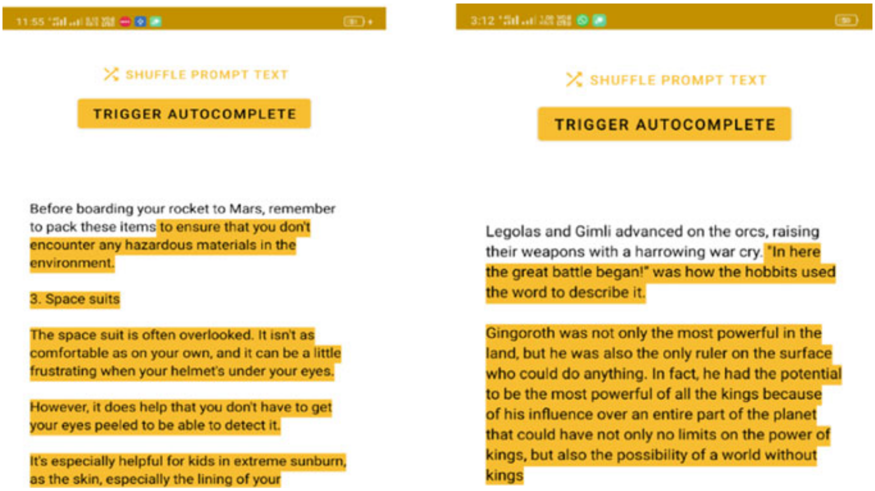


Fig. 10 Highlighted mark shows the predicted text

Table 1 Performance task

Model	SQuAD (EM/F1)
BERT-base	81.2/88.5
DistilBERT	77.7/85.8
DistilBERT (D)	79.1/86.9

performance and by fine-tuning the parameters in every layer. The model is achieving the F1 score 85.8 and EM score of 77.7 on the test set. Based on error analysis, there is a gap between the answer and has no answer prediction with the model performance, we can achieve in the later stage and we successfully predict the answers with matching the ground truth of the result.

The text generation using GPT-2 which is powerful transformer model which is used to generate the text capabilities or predict the next sentence. We have successfully generated the text from a raw text input and predict the next sentence from it. The architecture produces the coherent and fluent in the writhing the word. By the GPT-2 model which consists of the different sizes of the model and trained on the WebText data set and performed the state of the art in generating the text with accuracy in the related word and speed of the generation of word is good. We found that the model is producing the long text with as natural as human word.

Future Scope:

- Finding and extracting the large business problems from the question answering system.
- Dynamic selection of the answer prediction based on the user intent.

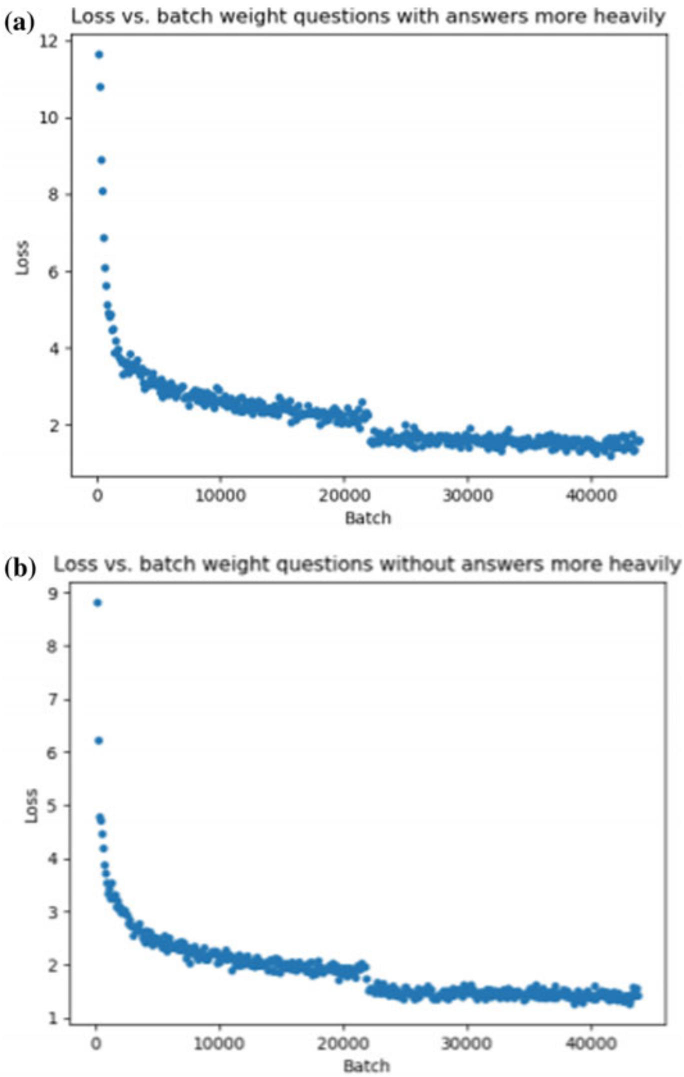


Fig. 11 Loss change with the training batches. **a** The question answers with heavy weights and **b** without more weights

- Automation in the finding the right solution or answer in the question and answer-related websites like Quora, Stack Overflow.
- Providing high accuracy in finding the answer without providing long related document or article.
- Providing quick review of the entire text.
- Adding the speech recognition capacities in question answering system and enabling the reading recognition for the disability person.

References

1. Annamoradnejad I, Fazli M, Habibi J (2020) Predicting subjective features from questions on QA websites using BERT. In: 2020 6th international conference on web research (ICWR)
2. Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805
3. Rajpurkar Z et al (2016) Lopyrev. Squad: 100,000+ questions for machine comprehension of text. arXiv:1606.05250
4. Radford A, Wu J, Child R, Luan D, Amadei D, Sutskever I (2019) Language models are unsupervised multitask learners. Technical report, OpenAI
5. Arustselvan M, Chellatamilan T, Vijaya Kumar M (2013) A statistical language modelling approach for question answering system. Int J Sci Res Publ 3(5). 1 ISSN 2250-3153
6. Medved' M, Horák A (2018) Sentence and word embedding employed in open question-answering. natural language processing centre, Faculty of Informatics, Masaryk University, Botanická 68a, 602 00, Brno, Czech Republic
7. Pennington J, Socher R, Manning CD (2014) GloVe: global vectors for word representation. Computer Science Department, Stanford University, Stanford, CA 94305
8. Wang S, Jiang J (2017) Machine comprehension using match-LSTM and answer pointer. Singapore Management University. Conference paper at ICLR, School of Information Systems
9. Yu L, Zhang W, Wang J, Yu Y SeqGAN: sequence generative adversarial nets with policy gradient. Shanghai Jiao Tong University, University College London. arXiv:1609.05473v6[cs.LG]. Accessed on 25 Aug 2017
10. Question Answering system framework: https://towardsdatascience.com/building-an-application-of-question-answering-system-from-scratch-2dfc53f760aa
11. Mathur A, Haider MTU (2015) Question answering system: a survey. In: 2015 international conference on smart technologies and management for computing communication controls energy and materials (ICSTM)
12. Vaswani A, Brain G, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, Kaiser Ł, Polosukhin Attention is all you need. https://arxiv.org/pdf/1706.03762v5 [cs.CL] 6 Dec 2017
13. Schwager S, Solitario J Question and answering on SQuAD 2.0: BERT is all you need. https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/reports/default/15812785.pdf
14. GPT-2 Architecture: http://humanssingularity.com/gpt2sampling/
15. Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding with unsupervised learning. Technical report, OpenAI
16. GPT2 training pipeline: http://jalammar.github.io/illustrated-gpt2/
17. Sanh V, Debut L, Chaumond J, Wolf T DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. https://arxiv.org/pdf/1910.01108.pdf

Improved LeNet Model for Flower Classification Using GPU Computing

Ritika Bhardwaj, Muskan Gupta, Akshita Pathania, and J. Sairabanu

Abstract Convolutional neural networks can be used for recognizing and categorizing wide range of image datasets. In the paper, we have proposed an improved LeNet model for flower classification. This revised LeNet is an improved version of modified LeNet model in paper (Mitrovic and Milosevic, Flower classification with convolution neural network). This model can be used for recognizing and classifying five different types of flowers. The dataset (flower recognition) is taken from Kaggle. CNN algorithms require a lot of mathematical convolutional computations which take a lot of time and has high computational demand. GPUs can be used for accelerating image processing. Also, we can parallelize CPU cores to process images faster. Along with the implementation of the improved LeNet model, this paper also differentiates between the performance of parallelized CPUs and a GPU. In the paper, we will be using Ray library in Python for parallelizing the multicore CPUs.

Keywords Convolution neural network (CNN) · LeNet · Flower classification · GPU · Sequential processing · Parallel processing · Ray library · Improved LeNet

1 Introduction

In recent years, plenty of research has been conducted for image identification, NLP and such other fields using neural networks. CNN has proved to be helpful in extensive range of applications. It is used to recognize patterns in images. As a result of its popularity over time, different versions of CNN are developed to increase accuracy and decrease the time required for its processing. CNN are widely used for image classification mostly on image datasets like CIFAR-10 [1].

One of the versions of CNN-LeNet model was introduced [2] to be applied for recognizing characters. Later, the study [3] was applied LeNet for classification of

R. Bhardwaj (✉) · M. Gupta · A. Pathania · J. Sairabanu
School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu 632014, India
e-mail: bhardwaj.ritika2000@gmail.com

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_9

Alzheimer's disease. After that, paper [4] provided a comparison of different CNN models like LeNet, GoogLeNet, etc. These models were used for the classification of medical pictures of anatomy.

The goal of our paper is to optimize the present model in such a way that it can accurately categorize different types of the flower on giving the image with an improved accuracy and processing speed than the present system.

In flower classification, there are a lot of similarities between flower classes as compared to dogs or cars. As flowers are non-rigid, we can see a lot of variation in flowers even within classes as they can deform in many ways [5].

Many papers and research have been done for classification of flowers. Software that can categorize flower type given an image can accord to research fields like horticulture and botany. The paper [6] has included 'shape' as an important component. They have built a visual vocabulary highlighting various facets which help to differentiate between flowers. In paper [7], a LeNet model with the Sigmoid Uniform function is proposed as modified LeNet, which gives better accuracy between AlexNet and LeNet, which is 73% for flower classification. However, the paper [7] has considered only three types of flowers. The result from the evaluation and analysis of these models (discussed above) enunciates the need of an improved architecture because the above research is not much concerned about the processing time. Data augmentation plays an important role in image processing and classification. Techniques like cropping, rotating, and flipping images are used for getting better results [8].

GPUs are widely used in training deep learning models because they can offer high bandwidth. GPU helps in implementing parallelization and is superior to CPU in many ways [9]. GPUs do not have a lot of memory capacity. GPUs can do the mathematical convolutional operations much faster than CPUs and can gain speedup up to 60% more than CPUs. Graphics processing unit (GPU) and parallelizing multiple CPU cores can be used as different ways for processing the images faster [10]. CNN on a large datasets become computationally intensive. Hence, CNN can be used along with parallel implementations in GPU [11]. In this paper, flower recognition image dataset is utilized available on Kaggle.

In this paper, we aim to propose a new version of modified LeNet [7] applied to CPU, GPU and multiple CPU processing system at different epochs. For implementing parallelized CPUs and GPU core, Python's Ray library is utilized. Ray can be used for developing emerging AI applications like image categorization, face recognition, etc. Ray library can identify the number of cores of CPU and GPU of a system. Parallelizing multiple cores of CPU results in the speedup of the CNN network significantly [12]. Our main objective is to present a more accurate model which can categorize flower classes based on the input image.

The paper is organized as follows. Section 2 has a description section about CNN and LeNet architecture. This is followed by a short description about the dataset. The existing method and proposed methodology are elaborated in fourth and fifth section, respectively. A comparison is drawn from the results and consequently conclusions are made in last section.

2 Convolutional Neural Networks

CNN networks are trained by firstly feeding a set of flower images along with their labels. These images are then passed through a heap of layers like the convolutional layer, pooling, Dropout and fully connected layers.

2.1 Convolutional Layer

This is first layer in CNN architectures. It helps to point out the features from the images of the dataset. This layer constitutes the convolution matrix operations using filters. These operations require a lot of processing power. Hence, in this project, we use GPU for handling complex computational processes. Filter is put over the image and then dot product is done between filter and image.

2.2 Pooling Layer

After convolutional layer, pooling layer is applied. The layer aims to decrease computational cost by reducing the dimensions of the convolved feature map. This is carried out by reducing the connections between the layers and it works on each feature map independently. There are various types of pooling operations like max or average pooling. This results in generating the images with compact dimensionalities that simplify the work for following layer.

Various convolutions and then pooling layers are applied to the model, then comes the output layers. This constitutes the flattening layer and fully connected layers with activation functions.

2.3 Flattening

Flattening is used for transforming the output data from previous convolutional layers into one-dimensional array. The fully connected layers intakes only one-dimensional data.

2.4 Fully Connected Layer

The fully connected (FC) layer attaches the neurons of different previous layers. These layers shape the CNN architecture and are placed before the output layer. In

FC layers, the product of weights and the previous layer are calculated. This shows the association between features of input image and of output category.

2.5 Dropout

Dropout layer is used to prevent the overfitting caused by attaching features to fully connected layer. When a model works too good for a training dataset impacting the model's functioning negatively when being applied on a test data. In dropout layer, neurons are dropped randomly to prevent overfitting.

2.6 LeNet-5 CNN Architecture

The LeNet-5 architecture can be seen in Fig. 1. Yann LeCun et al. presented it in 1989. It is one of the youngest and the most basic CNN architecture. It constitutes seven layers—three convolutional layers followed by two pooling layers and two FC layers.

The first layer is convolutional layer that intakes an input image of dimensions 32 × 32. The layer convolves the 32 × 32 input image using six filters resulting in 28 × 28 × 6 dimension image. This is connected to a pooling layer resulting in 14 × 14 × 6 feature map. This layer applies a filter of size (2, 2) and stride of 2. Layer 3 is again a convolutional layer with 5 × 5 convolutional kernels and 16 filters. Layer 4 is again a pooling layer resulting in a 10 × 10 × 16 feature map using filter of size of 2 × 2. Layer 5 is convolution layer with 120 filters of size (5, 5). Fourth layer (pooling layer) and fifth layer (convolutional layer) are completely connected. The

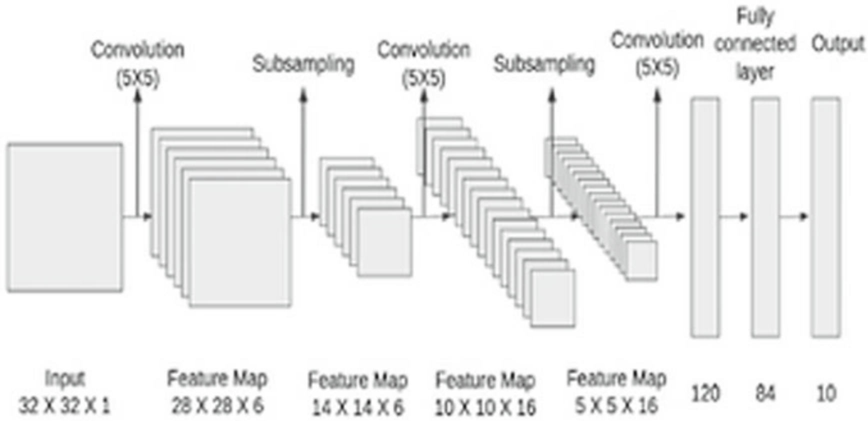


Fig. 1 LeNet architecture

Table 1 Flower types

Flower type	Number of images
Daisy	769
Sunflower	734
Rose	784
Tulip	984
Dandelion	1055

5th layer is labelled as a convolutional layer and not FC. This is done because in some cases, the size of output can be more than 1 × 1. Layer 6 is fully connected to layer 5. It results in 84 feature graphs as output. The last layer is the output layer resulting in ten neurons (output variables).

3 Dataset

3.1 Dataset Description

The paper uses a flower classification dataset [13] available on Kaggle. Alexander Mamaev has gathered and uploaded the dataset on Kaggle. It has 4242 images of various flower categories. The flower images in the dataset belong to categories: daisy, dandelion, rose, sunflower and tulip. Table 1 gives the categories and number of images in each category of the dataset.

Each of the photos under above mentioned categories is in the jpg format. The images have different height and weight; hence, during image pre-processing they are resized to 100 × 100 × 3. The dataset contains a wide variety of images of five types of flowers, so can be seen as a real-world example.

3.2 Dataset Pre-processing

The images are converted to a numeric array and resized to 100 × 100 × 3. The train test split is 75–25%. Image augmentation can be used for expanding the size of the dataset. With the help of image augmentation, we can get new augmented photos from the original dataset. On using the augmentation techniques like shifting or flipping, the target class remains unchanged.

For the flower dataset, image augmentation applied using the ImageDataGenerator() from Keras. The rotation range applied is 20 which will randomly rotate images through 20 degrees. The parameter height_shift_range is equal to 0.2 for shifting images vertically. The width_shift range is equal to 0.2 for shifting horizontally. Zoom range is also equal to 0.2.

The image augmentation techniques inculcate certain variation in data images. This allows our CNN model to generalize on new images. This results in a more robust model when trained on new and altered images.

4 Existing Approach

In the paper [7], modified LeNet model is implemented. It utilizes the sigmoid uniform function for initializing the weight and bias. The LeNet model uses Nesterov's momentum with value set to 0.9. Decay rate is set to 0.1 and step as 100,000 in the gradient descent method for optimization. L2 normalization is used for scaling the gradients to prevent exploding the gradients. Learning rate (lr) is initialized as 0.0001.

The model basically included two convolutional layers and each followed by a pooling layer. The initial convolutional layer is used to filter the pictures with 50 filters of size (5, 5) and stride (1, 1). The next convolutional layer comprises 100 filters. Both convolutional layers are connected to max pooling layers. Pooling layers has pool size (2, 2) and stride 1. Then there are two FC dense layers. This model utilizes ReLU and SoftMax activation function. The modified LeNet achieved an accuracy of 73.41%. Also, the model considers only three types of flower classification.

5 Improved LeNet Model

In the paper, we present an improved LeNet CNN model. Improvements have been made to the modified LeNet model [7]. The modifications include using an Adam optimizer with decay = INIT_LR/EPOCHS instead of SGD. Hence, nesterov's momentum is not being utilized. The sigmoid function is used as kernel initializer. Learning rate is initialized to 0.0005.

Keras is used for implementing the improved LeNet model on Kaggle. It acts as an interface for TensorFlow for ANN.

In Fig. 2, we can see the improved LeNet model. This model has two convolutional layers and two max pooling layers (As in LeNet-5). The first layer intakes the $100 \times 100 \times 3$ input picture with 20 filters of size 5×5 and stride of 1. This is followed by max pooling layer of poolsize (2, 2) and stride (2, 2). Second convolutional layer has 50 kernels of size 5×5. This is followed by second max pooling layer. Then, there is a dense layer with 500 units. It is included in the output layer. ReLU activation is used in the convolutional and first FC layer. SoftMax activation used in last fully connected layer to output layer.

Ray library is used for GPU computing. Using ray.init(), Ray can utilize all cores of the machine. To configure the number of cores, mention it as a parameter in init function—ray.init(num_gpus = 1). Ray allows arbitrary functions that can be

Layer (type)	Output Shape	Param #
conv2d_2 (Conv2D)	(None, 96, 96, 20)	1520
max_pooling2d_2 (MaxPooling2	(None, 48, 48, 20)	0
conv2d_3 (Conv2D)	(None, 44, 44, 50)	25050
max_pooling2d_3 (MaxPooling2	(None, 22, 22, 50)	0
flatten_1 (Flatten)	(None, 24200)	0
dense (Dense)	(None, 500)	12100500
dense_1 (Dense)	(None, 5)	2505

Total params: 12,129,575
Trainable params: 12,129,575
Non-trainable params: 0

Fig. 2 Improved LeNet

executed asynchronously. These asynchronous Ray functions are called remote functions. Using @ray.remote() decorator, a regular Python function will become a ray remote function. For invoking this remote function, remote method will be used. It returns an object reference and then creates a task that will be executed on a worker process. Remote operations are asynchronous. Ray.get() will return an object or a list of objects. This is synchronous operation. For utilizing GPU computing, initialize ray.init() specifying the gpus in parameter—ray.init(num_gpus = 1). For parallelizing multiple CPU cores, we can initialize ray function with 4 CPU cores—ray.init(num_cpus = 4). We need to include the GPU or CPU core in @ray.remote() decorator.

The proposed improved LeNet model has extended the modified LeNet [7] model to all five flower types. The modified LeNet [7] is applied only to dandelion, sunflower and tulip. While the proposed improved model is applied to all five types (dandelion, tulip, sunflower, rose and daisy).

6 Results

For each CNN model, the evaluation metrics including accuracy, precision, recall and F1-score. Confusion matrix is also included in the result. Accuracy defines the percentage of correct predictions.

Figure 3 shows the results for LeNet model. Accuracy is nearly 57%. Other evaluation metrics are also unsatisfactory. Some of the records are correctly classified especially in sunflower class. But the tulip images are falsely classified as sunflower and dandelion as tulip.

Figure 4 shows the evaluation metrics of the modified LeNet model. Modified LeNet has an accuracy of 73%. There is considerable increase in accuracy from

```
========================Evaluation Metrics========================
 # of classes:    3
 Accuracy:        0.5773
 Precision:       0.5837
 Recall:          0.5777
 F1 Score:        0.5704
Precision, recall & F1: macro-averaged (equally weighted avg. of 3 classes)


=========================Confusion Matrix=========================
   0   1   2
-------------
  65  39  43 | 0 = dandelion
  12 111  23 | 1 = sunflower
  25  44  78 | 2 = tulip

Confusion matrix format: Actual (rowClass) predicted as (columnClass) N times
```

Fig. 3 LeNet model

```
========================Evaluation Metrics========================
 # of classes:    3
 Accuracy:        0.7341
 Precision:       0.7587
 Recall:          0.7341
 F1 Score:        0.7346
Precision, recall & F1: macro-averaged (equally weighted avg. of 3 classes)


=========================Confusion Matrix=========================
   0   1   2
-------------
 123  21   3 | 0 = dandelion
  33 107   6 | 1 = sunflower
  29  25  93 | 2 = tulip

Confusion matrix format: Actual (rowClass) predicted as (columnClass) N times
```

Fig. 4 Modified LeNet model

LeNet model. But still sunflower images are wrongly recognized as dandelion and dandelion as sunflower.

For the proposed improved LeNet model, the evaluation metrics, accuracy and confusion matrix can be seen in Fig. 5. A classification report is also shown. The model has better accuracy than all other. Additionally, the model can classify flowers into five different types. Unlike the modified LeNet model [7], the sunflower images are not wrongly classified. Other metrics like precision, F1-score and recall are also improved. This shows that our model works better with the dataset in identifying the categories of flowers.

Refer Figs. 6 and 7 to see the loss and accuracy of the improved LeNet. Model loss indicates prediction error of CNN. Graph shows how model loss changes with the number of epochs. As the epochs (iterations) increases, the loss decreases. The

====================CLASSIFICATION REPORT====================

	precision	recall	f1-score	support
dandelion	0.80	0.77	0.79	133
daisy	0.86	0.81	0.83	218
sunflower	0.74	0.76	0.75	174
tulip	0.85	0.90	0.88	167
rose	0.75	0.75	0.75	173
accuracy			0.80	865
macro avg	0.80	0.80	0.80	865
weighted avg	0.80	0.80	0.80	865

====================EVALUATION METRICS======================

```
# of classes  : 5
Precision     : 0.797816
Recall        : 0.798968
F1 score      : 0.798014

Train Accuracy: 91.20879173278809
Test Accuracy : 80.0000011920929
```

====================CONFUSION MATRIX========================

	dandelion	daisy	sunflower	tulip	rose
dandelion	103	13	7	6	4
daisy	14	177	6	14	7
sunflower	4	8	132	3	27
tulip	2	3	5	151	6
rose	6	5	29	4	129

Fig. 5 Improved LeNet model

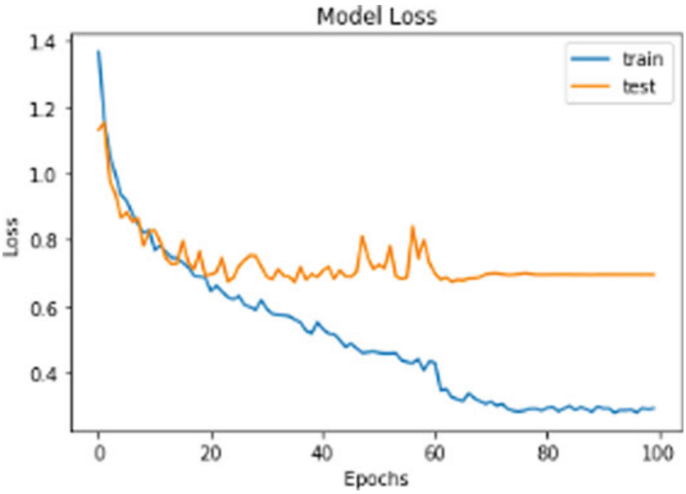


Fig. 6 Model loss

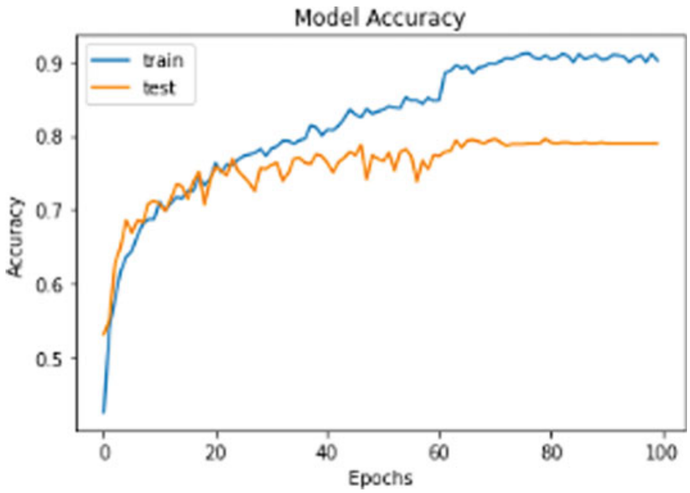


Fig. 7 Model accuracy

graph in Fig. 7 shows how the accuracy of the model increases with the increase in the number of epochs using GPU computing. As epochs increase, the model becomes better trained, and hence, the accuracy increases.

Figure 8 shows that how utilization of GPU for image processing can reduce the time taken for building, training and evaluating the network model. With the increase in epochs, the time also increases. But this increase in the case of parallelised CPUs and serial Python is very high as compared to GPU. Serial python takes most of the time for 100 epochs while 4 CPU takes considerably less time compared to

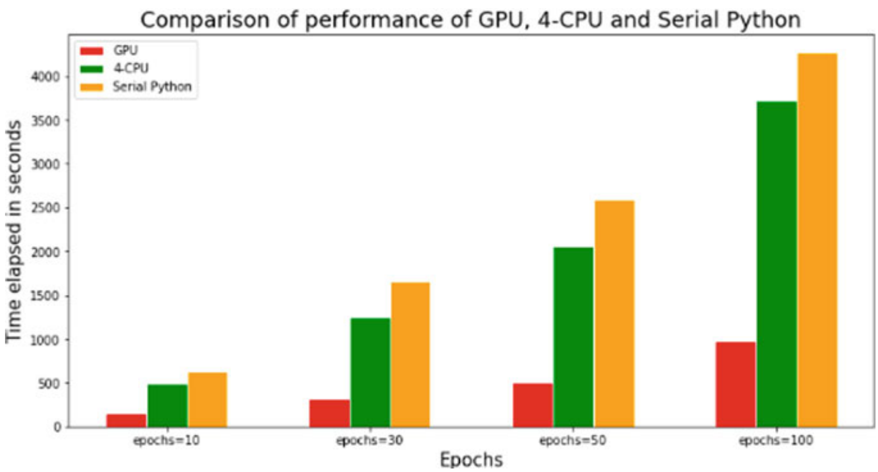


Fig. 8 Performance of GPU, 4 CPU and serial Python (1 CPU)

Table 2 Evaluation metrics of LeNet models

	LeNet	Mod LeNet (with 3 flower types)	Improved LeNet (with 5 flower types)
Recall	0.57	0.73	0.797
Precision	0.58	0.75	0.798
F1-score	0.57	0.73	0.798
Accuracy (%)	57	73	80

serial Python (1 CPU). Serial Python refers to 1 CPU. Using GPU accelerates image processing and thus performs the convolutional operations much faster resulting in less time for all the epochs.

Utilization of GPU not only decreases the time but also considerably increases accuracy as compared to parallel CPUs and serial python.

The evaluation metrics for the CNN models are shown in the Table 2. To conclude, the proposed improved LeNet model gives better and satisfying results in comparison with other models. Also, the proposed model has extended the mod LeNet to all five flower types.

7 Conclusion

We have successfully implemented an improved LeNet model with accuracy of 80%. The model can recognize the photos of various flowers and categorizes them with best accuracy. The dataset has five classes: daisy, dandelion, sunflower, tulip and rose. The model was trained on Kaggle using GPU computing for image processing.

The paper has introduced a CNN model that can handle diverse image dataset. The research done before this work on flower classification has photos zoomed on the flower in the centre.

LeNet CNN with a sigmoid function [7] for a dataset considering only three flower types has accuracy 73%. The results in the Table 2 show that the proposed improved LeNet model is extended to all five flower types and has better accuracy. When compared with previous works which are mostly based on recognizing alike pictures, the achieved accuracy is adequate.

4242 images of five different flowers are being processed by the network model. Hence, there is a lot of processing required which could take a lot of time. Training the network model on GPU decreases the computation time drastically and improves the accuracy. Since GPU computing is being used, the accuracy of the model shifts between 79 and 82%.

As seen in Table 3, GPU computing has accelerated image processing. The usage of four parallel CPUs is better than serial Python. GPU implementation results in significant speedup compared to CPUs. The CNN architecture involves a lot of mathematical operations. GPU can perform these complex matrix calculations very

Table 3 Time taken by the LeNet model for different epochs and different processors

Processor	Epochs	Time taken for improved LeNet (s)
GPU	10	153
4 CPU cores	10	488
Serial python	10	624.49
GPU	30	308.80
4 CPU cores	30	1240.736
Serial python	30	1662.62
GPU	50	501.83
4 CPU cores	50	2057.966
Serial python	50	2592.12
GPU	100	976.39
4 CPU cores	100	3724.56
Serial python	100	4067.052

fast. Hence, GPU is used to speed up image processing. The latency is overridden by the number of small cores that are present in GPU.

References

1. Chauhan R, Kumar K, Joshi GRC (2018) Convolutional neural network (CNN) for image detection and recognition
2. LeCun Y, Bottu BL, Bengio Y, Haffner P (1998) Gradient based learning applied to document recognition. In: Proceedings of the IEEE, p 86
3. Sarraf S, Ghassem T (2016) Classification of Alzheimer's disease using FMRI data and deep learning convolution neural networks. arXiv:1603.08631
4. Khan S, Suet-Peng Y (2017) A deep learning architecture for classifying medical images of anatomy object. In: Asia–Pacific signal and information processing association annual summit and conference (APSIPA ASC). IEEE, pp 1661–1668
5. Nilsback M-E, Zisserman A (2008) Automated flower classification over a large number of classes. In: Sixth Indian conference on computer vision, graphics & image processing. IEEE, pp 722–729
6. Siraj F, Salahuddin MA, Yusof SAM (2010) Digital image classification for Malaysian blooming flower. In: Second international conference on computational intelligence modelling and simulation. IEEE, pp 33–38
7. Mitrovic K, Milosevic D (2019) Flower classification with convolution neural network
8. Wang J, Perez L (2017) The effectiveness of data augmentation in image classification using deep learning, Stanford University
9. Syberfeldt A, Ekblom T (2017) A comparative evaluation of the GPU versus the CPU for parallelization of evolutionary algorithms through multiple independent runs. Int J Comput Sci Inform Technol (IJCSIT) 9(3)
10. Memon ZA, Samad F, Awan ZR, Aziz A (2017) CPU-GPU processing, September 2017
11. Li X, Zhang G, Howie Huang H, Wang Z, Zheng W (2016) Performance analysis of GPU-based convolutional neural networks. In: 45th international conference on parallel processing. IEEE. https://doi.org/10.1109/ICPP.2016.15

12. Moritz P, Nishihara R, Wang S, Tumanov A, Liaw R, Liang E, Elibol M, Yang Z, Paul W, Jordan MI, Stoica I (2018) Ray: a distributed framework for emerging AI applications. In: Proceedings of the 13th USENIX symposium on operating systems design and implementation(OSDI'18)
13. Mamaev A (2018) Flowers recognition [Online]. Available: https://www.kaggle.com/alxmamaev/flowers-recognition/home

Distributed Computing Meets Movable Wireless Communications in Next Generation Mobile Communication Networks (NGMCN)

A. Madhuri, S. Sindhura, D. Swapna, S. Phani Praveen, and T. Sri Lakshmi

Abstract Distributed technology will have a significant influence on wireless communications on the move in next generation cellular networks. On the one hand, portable distributed computing (PDC) solutions are enabled by the integration of distributed computing into the mobile platform. The distributed powerful computing platforms for radio access networks, on the other hand, have given rise to a unique notion of distributed radio admittance web (D-RAW). In this paper, we look at how topologies are configured and rate allocation challenge in D-RAW with the goal of improving PDC customers' end-to-end performance in the future cellular networks of the next generation. To deal with delayed channel state information, we take a decision theory method. The findings of simulations suggest that future mobile wireless network design and operation may be simplified. Distributed computing has a big impact, and the suggested method is capable of reaching significant performance benefits in comparison with existing systems.

Keywords Generation cellular networks · Distributed radio access networks · Distributed computing

A. Madhuri (✉) · D. Swapna · S. Phani Praveen · T. Sri Lakshmi
Department of Computer Science & Engineering, PVP Siddhartha Institute of Technology, Vijayawada, India
e-mail: madhuria@pvpsiddhartha.ac.in

D. Swapna
e-mail: dswapna@pvpsiddhartha.ac.in

T. Sri Lakshmi
e-mail: tslakshmi@pvpsiddhartha.ac.in

S. Sindhura
Department of CSE, NRI Institute of Technology, Agiripalli, India

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_10

1 Introduction

Disseminated processing has arrived at maybe the most sweltering point in both the scholarly local area and industry. Conveyed figuring is a model for empowering on-request admittance to a typical configurable asset arrangement (e.g., laborer's, stockpiling, applications, organizations, and so forth) asset save, quick adaptability, and essential organization [1]. Some assistance models are upheld, including distributed planning as an aide, the distributed stage as an aide, and the distributed system as an administrator. Circulated processing has been the subject of essential thought, and some corporate organizations, like Google App Engine, Microsoft Azure, and, have offered kinds of client service.

Dispersed registering will altogether influence the arranging and activity of cutting-edge compact far-off PDA associations. According to one perspective, with propels in correspondence and flexible far-off gadgets, increasingly more end clients are getting to disseminated preparing systems through cell phones, like PDAs and tablets. The versatile climate distributed computing reconciliation empowers portable distributed computing (PDC), which is by and large viewed as a promising perspective of convenient registering with a tremendous market [2]. PDC allows you to offload the preparing force, and capacity needs of cell phones into the inconceivable distributed logging stages, conquering any hindrance between growing processing requests and the standard, convenient registering progressions with restricted logging assets, stockpiling, and energy in cell phones.

Furthermore, the inconceivable development of circulated figuring may likewise be valuable for radio admittance web (RAW) (despite versatile end clients), recommending a novel thought of radio admittance web in the Distributed (D-RAW) [3, 4]. In contrast to the present cell associations, where baseband status recording assets are situated at every single telephone spot, in D-RAW, figure assets are situated in a distant central association of distributed using a staggering measure of assets estimation. This move from a summed up system to a typical structure for baseband planning can have essential advantages: work cost investment funds through incorporated upkeep; permit a superior change of charges; improve the execution of the association because of the techniques of sign administration worked with of last age; decline in energy utilization because of maltreatment of assortments in batteries.

While some incredible work was never appropriated to both end customers and access affiliations, these two critical areas were overall managed openly when creating. D-RAW's joint assessment on PDC for the state-of-the-art cell networks has not been the subject of past effort. In this article, we will look at the geographic plan, and rate parcel issue in D-RAW resolved to revive beginning to end Connection Oriented Protocol execution of PDC clients in driving edge cell affiliations.

In D-RAW, the most significant issues are caused by the fact that the channel status data (CSI) is jumbled up as a result of the delay in receiving and sending such data. A wrecked CSI impacts the entire D-RAW, yet moreover a distant relationship as a rule. Since it is difficult to determine this issue using hypothetical customary data strategy [5], we embrace a speculative philosophy of choice, which has made systems adjust with the effects of fierce and conceded CSIs [6]. An ideal technique subject to the specific advancement of the geographic environment and cost spread can be found. The reenactment results show that the course of action and development of future helpful distant affiliations can be ultimately impacted by appropriate taking care of. The proposed conspiracy is prepared toward achieving essential headway in execution on top of existing plans.

2 Related Work

Figure 1 shows a state-of-the-art cell network with D-RAW and PDC redesigns. Shown and customary mobile stations (MSs), MSs in D-RAW are modified considering the path that by a long shot, the more significant part of sign organization and components occurs in the far away affiliation's distributed. The MSs are connected with the far-off affiliation's distributed through backhaul networks. To all the more possible arrange information improvements start to finish, a standard blueprint is confined to Connection Oriented Protocol. Plan sharp, and we can have an alternate Connection Oriented Protocol vendor on the edge of the distant affiliation's distributed. Split Connection Oriented Protocol delegate can be acted in the change from structure plan improvement (SPI-GW) to LTE framework as customer information streams are moved to SAE-GW preceding being directed from the World Wide Web [7]. It sees each segment and a short time later store and advances the part in the second Connection Oriented Protocol alliance. The distributed application provider does not have to change its framework for PDC customers by picking the split Connection Oriented Protocol mediator.

Figure 2 shows the steady connection between cell phones, distant association distributed, and back-end laborers. The Connection Oriented Protocol streams sent by the flexible distributed organizations are conveyed from cell phones to the back-end laborers in the distributed.

The portable client and the back-end worker into two associations and supports a constant association among itself and the back-end worker. In the interim, the remote organization distributed conducts dynamic procedures on small organizations to offer

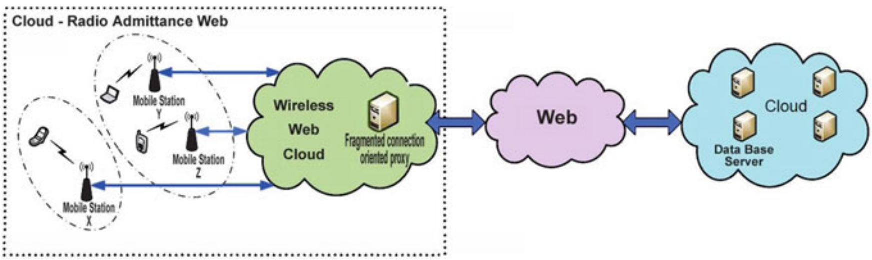


Fig. 1 Access network of PDC environment in the distributed radio

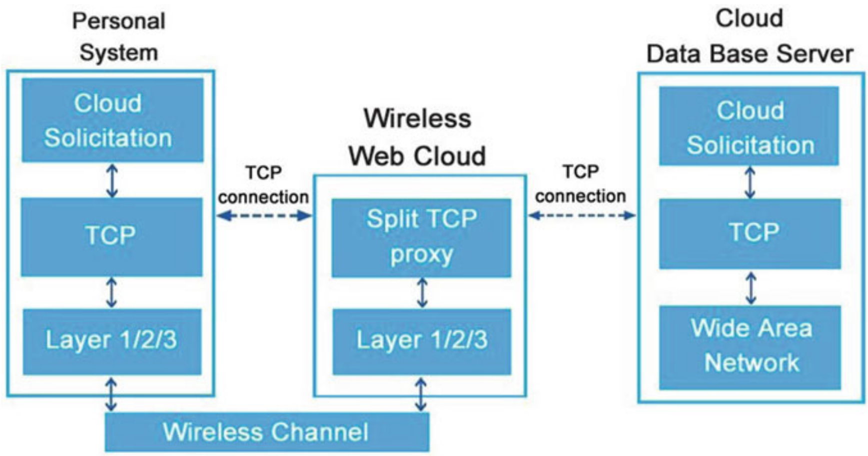


Fig. 2 Logical protocol data center, split Connection Oriented Protocol proxy, and mobile user. Stacks of network entities

the best assistance for the upper layer. Such unique activities incorporate geography arrangement and rate portion. Geography arrangement controls how the MSs help out one another. For example, in Fig. 3, MSs B and C structure a group to assist the two versatile clients combined while MS an itself is an additional bunch. In wake of grouping, the remote organization distributed needs to choose the information rates the versatile clients can send.

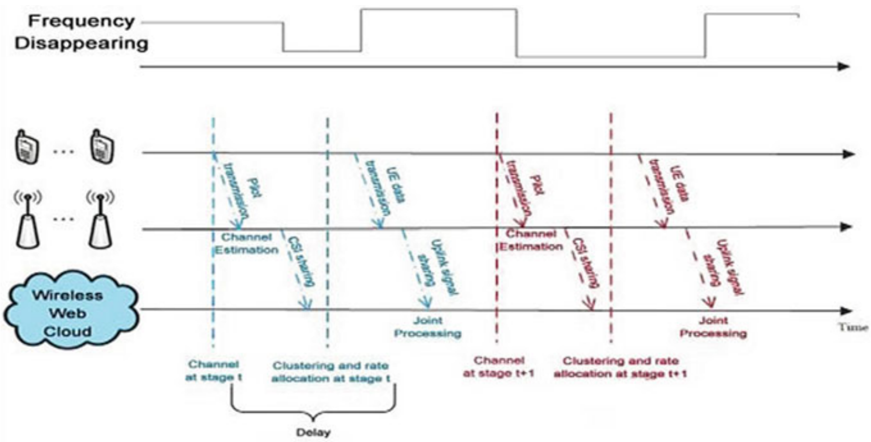


Fig. 3 Introduction of delay by networks backhaul in D-RAW

3 D-RAW with Afternoon Channel Status Information

If all else fails, backhaul networks in D-RAW are confined and have slack in CSI correspondence. This limit gives a deferral between the circumstance, the undeniable channel, and the natural channel. To perceive how the redirection enters D-RAW, we reflect a D-RAW showed in Fig. 1, and its resulting endeavors showed in Fig. 3. The CSI is gotten through the pilot signs got from the MS. After the channel is laid out, the CSI will be sent through the backhaul relationship to the distant affiliation's distributed. In the far away coalition distributed, you pick how to work with the MSs and the rates at which flexible clients can permit happening to secure CSIs. Then, at that point, customer information is passed on. Indistinct from CSI's appraisal and augmentation, client signals are permitted by versatile clients to the MS, then, at that point, made in the backhaul networks.

During this time of sections, open CSI is old. We can isolate the novel ability between the guaranteed state of the channel during this depiction of parts and that of data as a solitary number. We may provide the reference problem to the channel it represents by using the Markov chain channel model, which has been employed in all of the cases examined. In the Markov chains, we map the deferral in seconds over the course of extended periods of progress. Given the delay of progress, the conviction status can be settled, which is the legitimate past progress and data history level. A conviction state but close to the start of program t is a good state-space course. At present, the relative part's probability gives the state the errand t in bt, deduced as b (st). With structures, like time steps, we can know the degree of deferral of steps d.

With such weakness, understanding is only the affirmed conceded by pushes. So by then, at that point, we can get the concede conviction state and the change probability connection.

4 Proposed Methodology

In this portion, we study the introduction of Connection Oriented Protocol over D-RAW in PDC frameworks and study client reaction torpidity. Extending Connection Oriented Protocol throughput with the focal issue of reaction idleness is seen as an obliged stochastic improvement issue. We could form the start show that the covetous strategy (i.e., brief execution improvement) is ideal for addressing the stochastic protect issue. So by then, we pick the conviction state, which is the colossal reliability of the guaranteed picture states. The stochastic improvement issue for quick execution acquires changes into a deterministic discrete movement issue with the passed-on conviction.

4.1 Round Trip Time and Split Connection Oriented Protocol Throughput

Fragmented—Connection Oriented Protocol has become a remarkable show for moving information for master residence affiliations and inheritance cell affiliations. We expect that it should be a different option from colossal work in PDC's top-level cell affiliations. In this sense, in this report, we get detached Connection Oriented Protocol as the underlined transport layer show. In [8], an all-around robust Connection Oriented Protocol execution model is made. It has been employed in cross-layer applications to cultivate Connection Oriented Protocol execution (e.g., [9]). In this piece, we discharge up the past work to consider yielded CSI in the Connection Oriented Protocol execution prototypical.

It should not something be said about we look at RTT first? Figure 4 shows the total circle times for the strong distributed relationship in D-RAW. There are two kinds of RTT. RTT1 figures out RTT among clients and the split Connection Oriented Protocol ace at the far-off affiliation's distributed edge. RTT2 is just the RTT between the supervising Connection Oriented Protocol arranged capably and the distributed back-end trained professional. D-RAW influences RTT1. RTT1 Wireless and backhaul uninhibitedly course the entire transmission plan through far away affiliations and backhaul.

Wireless T can be received from the fixed layer and the genuine layer models. The value of RTT2 can be evaluated using the methodology, for instance, time steps. We recognize that the most unbelievable irregularity size (MSS) is set, so a particular

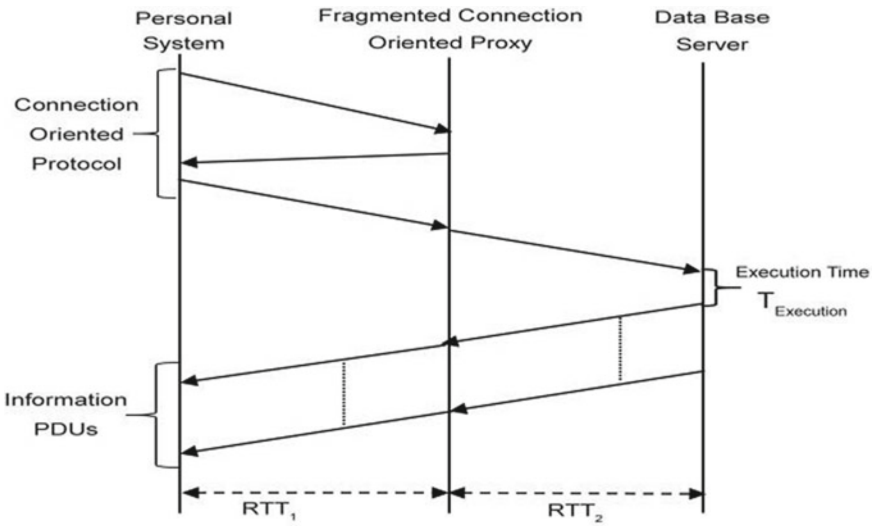


Fig. 4 Split Connection Oriented Protocol round trip times

segment sorts out some way to enter a single affiliation level profile without loss of arrangement.

On the off chance that a special Connection Oriented Protocol message cannot be stacked on edge, the size of the fix layer profile may be fundamental to the choice [10]. Line postpone is not covered. Since we revolve around yielded CSI in this document, we can, by and large, consider the deferral in the line as a consistent commitment to RTT. To the extent of execution, the deferral of the line can be approximated by its quantifiable normal. This assumption looks reasonable since the time estimation of the things in the line is more noticeable than that of the far-off channel, as there is much information to confer. If selections are relatively few, line deferral can be altogether ignored. Therefore, we consider the line slack to be steady, which was considered in T backhaul.

Concerning the probability of dissatisfaction of the Connection Oriented Protocol bundle in faraway affiliations, the authentic level error likelihood is picked by the authentic level error likelihood, that is, the blackout probability and the HARQ affiliation layer show. A shut development joint is openly recorded as a printed version since a specific HARQ conspiracy is used.

L. Response dormancy per client

Reaction dormant is fundamental for adaptable distributed associations [11]. Since significant cycle designations are shown on specialist farms, PDC frameworks bear the adverse outcomes of reaction time achieved by the time course of action and exchanges between network segments. The inaction of the control is essentially directed by the gathering and the work structure, which is not the point of convergence of our work. Besides, as the going with bits will show, planning with inactivity can be improved through sensible D-RAW masterminding and activity.

M. Boosting Connection Oriented Protocol throughput with delayed CSI for mobile.

Distributed services

Because of the obsolete CSI, the introduction of versatile distributed relationship over split Connection Oriented Protocol is a discretionary variable influenced by D-RAW choices. We get a speculative decision framework, which has made designs to manage the rambunctious and conceded CSI impacts.

At the dispatch of program t, the bundling state St is an undetected offbeat variable. The distant association distributed chooses the collaborating MSs and offers flexible customers, displayed as in.

The starting to the farthest uttermost spans of the show of a versatile customer u is known as an unusual variable u (at, St), and $\Sigma B\ \eta u$ (at, St) is the finished return of the painting. Show the extent of time tasks considered h, alluded to as the skyline all out recorded as a printed duplicate of the Markov decision cycle [12]. The full-scale changes in the h skyline are $\Sigma t = h\ \Sigma u = B\ \eta u$ (at, St).

Also, we mean the inertia of the response portrayed in Eq. 1 as Tu, and power the lethargy to be under an α limit. To build the full-scale throughput of the Connection Oriented Protocol and cutoff the dormancy of the response, we have been ongoing with this legitimization issue.

$$\underset{a^t, t=1,2,\ldots,h}{\text{maximize}}\ \mathbb{R}\left[\frac{1}{h}\sum_{t=1}^{t=h}\sum_{u=1}^{u=B}\eta_u(a^t, S^t)\right] \tag{1}$$
$$\text{subject to } \mathbb{R}[\tau_u(a^t, S^t) < a], u = 1, \ldots, B, t = 1, \ldots, h.$$

The issue in (1) is an obliged stochastic structure-up issue. From the start, we propose an exciting framework, wherein the standard extent of the target work that can be created in the current opening undertaking is improved. Constantly end, it is ideal for the stochastic improvement issue (1) when $h = 1$. We can show that a positive way of thinking is ideal when we think about different horizons. The demanding project is according to the going with. Consider the horizon $h = 1$the best move to perform is the maximize of $E\ u = B\ \eta u\ (a1, S1)$, which is the activity given by the voracious outlook to assemble standard differentiations in a single development. Expect a horizon h, $h \geq 1$, and the best approach is the ravenous technique. By the greatness of the $h + 1$ horizon, given the theory that the best strategy cultivates the standard h steps, the improvement to foster the standard better charges is to expand the premium of $h+$ starting development, equivocal from the condition with the horizon 1. In this line, the rapacious focus is the best approach for the spilling over (1).

From the arrangement and deferral of the channel, we can get the conviction state but, which is the gigantic likelihood of the current CSI. The issue of stochastic movement in all those approaches can be changed to a deterministic improvement issue since the conviction states can choose the traditional worth of the yield and the inaction of the reaction. Similarly, the deterministic improvement issue is a numerical programming issue since the activities are discrete. Frameworks have been made to manage these number programming issues.

For example, medium-sized issues can be advantageously managed by the growing and ricocheting framework, and the extraordinarily massive number of programming can be settled using heuristics like inherent math.

5 Experimental Results

The introduction of the proposed stunt is depicted through NS2-subordinate PC redirections. We perform redirections using going with conditions. The teleportation occasion is 2.1 GHz. There are three MSs in the D-RAW. The strangest size of a collaborating pack is 2. The far-off channel is the Rayleigh faint channel; whose standardized Doppler shift goes from 0.01 to 0.06. The data transmission is 45 kHz. The union layer licenses you to give the overall edges on various events. For Connection Oriented Protocol streams, the payload size is 760 bytes. Wmax is 6 MSS. We ought to consider a web record application under the versatile distributed 2. Reaction torpid is 2 other PDC applications.

For instance, adaptable games and online social affiliations can, in like manner, be considered in our arrangement. Exceptionally the relationship of web crawlers coming from cells. The course of action of the dormancy major of the α reaction follows that recommended in [13], which is set to 0.35 s in the proposed stunt.

5.1 Execution Improvement

Surveyed the CSI delay in D-RAW utilizing the test unit. Figures 5 and 6 openly show execution and response inaction in low flexibility and high convenience conditions.

From these figures, we can see that the proposed design beats the current ones to the degree both edge total Connection Oriented Protocol throughput and response in action. In the low convenience circumstance, the outright Connection Oriented Protocol throughput of both the proposed interest and a stunning CSI flow strategy in D-RAW consistently diminishes as the deferral increments. By some coincidence, the proposed stunt accomplishes an ideal yield over the current strategy, for instance, regarding the CSI of ten models, by about 30%. In the meantime, for the current arrangement of a splendid CSI, the client response time increments as the CSI winds up being logically old. A preferred advancement over the proposed plan can be found in the high flexibility case displayed in Fig. 4.

Execution brilliant, the introduction of the current blueprint considering just the certified execution of the level is the most repulsive of the three. As to torpidity of response, if there should arise an occasion of low versatility, as the CSI develops, the

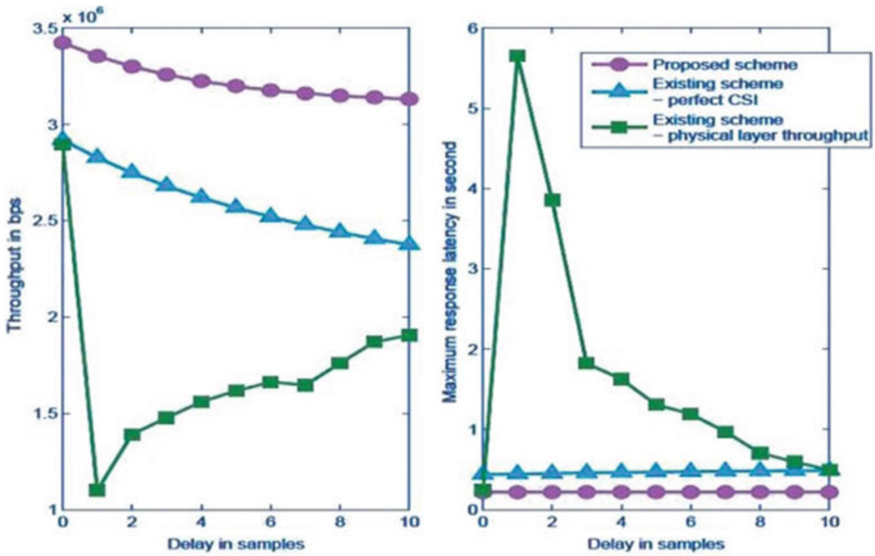


Fig. 5 (Low mobility case with normalized Doppler shift 0.01) The effect delayed CSI in D-RAW

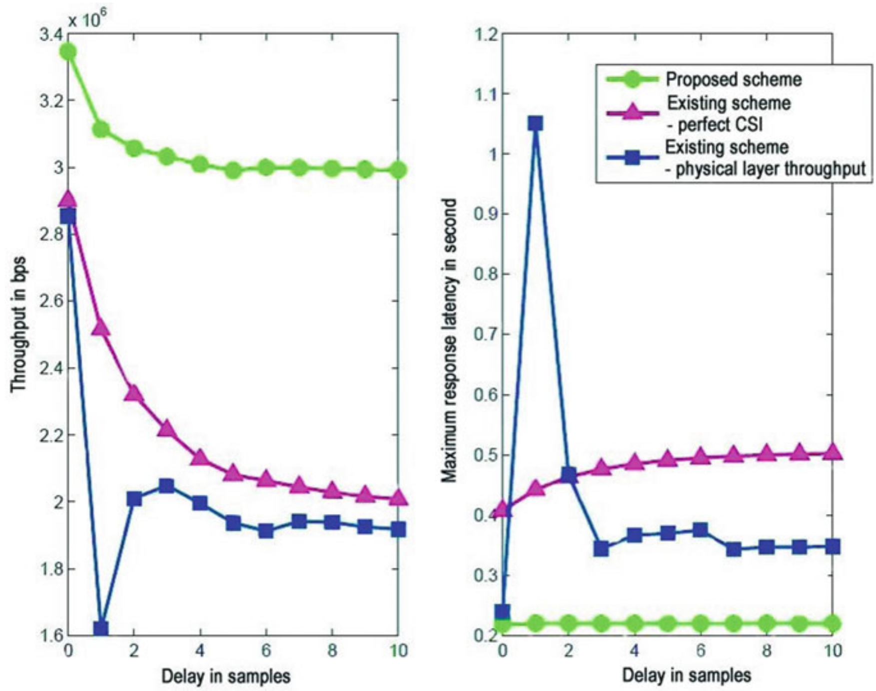


Fig. 6 (High mobility case with normalized Doppler shift0.06) The effect of delayed CSI in D-RAW

inaction is near that of the current undertaking that invites the magnificent CSI; for the condition of high adaptability, it outsmarts the current game plan by expecting an amazing CSI when the deferral is more essential than two models. As displayed in our past work [7], the current game plan that overhauls the authentic showcase of the level has the best execution than the current course of action that expects a wonderful CSI when the standard is the total musicality of a significant number of customers in the bundling. Likewise, its lead is the length of the deferral in CSI increases. Such an arrangement is not reasonable when the standard is the total Connection Oriented Protocol throughput of adaptable distributed affiliations. The characteristic clarification is that the authentic show does not just restrict the lead of Connection Oriented Protocol of the layer, yet what is more.

Full circle time and consistent quality start to finish. The method that develops layer execution finds a type of friendliness between blackout probability and speed task to achieve the most prominent layer execution, which is dangerous for adaptable distributed associations. So when the deferral is nearly nothing, say two models, the simple display diagram of the climbing level has the most dreadful show as the deferral broadens the feasibility of such a game plan to grow the real show of the level decreases.

This way, the presentation and latency of the Connection Oriented Protocol approach that of the current course of action by enduring bewildering CSIs. This is the explanation we can see a low CSI deferral locale top in these figures. From these revelations, we can see that it is essential to plan and work D-RAW to conform to Connection Oriented Protocol driving start to finish; moving adaptable distributed associations to best in class cell affiliations.

Note that RTT over landlines, RTT2, can impact both customer input idle and blueprint execution. With a colossal RTT2 or a little reaction torpidity edge, it may not be practical to find utilitarian reactions for (1). In the current conditions, various gadgets, for instance, claims control, should be used to confine the amount of PDC clients in the framework.

6 Conclusion

This article usually has investigated distributed RAN and versatile dispersed figuring in driving edge cell affiliations. In particular, the issue of the geographic plan and esteeming in D-RAW was examined to advance the beginning to end Connection Oriented Protocol execution of PDC customers in driving edge cell affiliations. We have proposed a speculative technique for choice to address the faulty CSI issue in D-RAW. The inactivity of the reaction experienced by each PDC client was shown as an obstacle. Using the re-foundation results, we have shown that our proposed plan can significantly improve the presentation of the design in regards to execution and inaction of PDC customer reactions. In particular, the deferred CSI in D-RAW generally influences receptiveness, and our proposed plot can diminish that impact, particularly in conditions of monster slack and high flexibility.

References

1. Pallis G (2010) Distributed computing: the new frontier of internet computing. IEEE Internet Comput 14(5):70–73
2. Phani Praveen S, Bala Murali Krishna T, Chawla SK, Anuradha CH (2021) Virtual private network flow detection in wireless sensor networks using machine learning techniques. Int J Sens Wirel Commun Control 11. https://doi.org/10.2174/2210327910666210104160027
3. Bhaumik SP, Chandrabose MK, Jataprolu G, Kumar A, Muralidhar P, Polakos V, Srinivasan, Woo T (2012) Distributed IQ: a framework for processing mobile stations in a data center. In: Proceedings ACM Mobicom'12 (Istanbul, Turkey)
4. China Mobile Research Institute, D-RAW: the road towards green RAN. Research Report, http://laMS.chinamobile.com/. Accessed on 18 July 2013
5. Goldsmith A, Effros M, Koetter R, Medard M, Ozdaglar A, Zheng L (2011) Beyond Shannon: the quest for fundamental performance limits of wireless adhoc networks. IEEE Comm Mag 49:195–205
6. Puterman M (1994) Markov decision processes: discrete stochastic dynamic programming. Wiley, Inc.
7. Sindhura S, Praveen SP, Syedbi S, Pratap VK, Krishna TBM (2021) An effective secure storage of data in cloud using ISSE encryption technique. Annals Roman Soc Cell Biol 5321–5329

8. Swapna D, Phani Praveen S (2020) An exploration of distributed access control mechanism using blockchain. In: Smart intelligent computing and applications. Springer, pp 13–20
9. Toledo A, Wang X, Lu B (2006) A cross-layer connection oriented protocol modelling framework for MIMO wireless systems. IEEE Trans Wirel Commun 5:920–929
10. Luo C, Yu FR, Ji H, Leung VC (2010) Cross-layer design for connection oriented protocol performance improvement in cognitive radio networks. IEEE Trans Veh Tech 59:2485–2495
11. Pathak A, Wang YA, Huang C, Greenberg A, Hu YC, Kern R, Li J, Ross KW (2010) Measuring and evaluating connection oriented protocol splitting for Distributed services. In: Proceedings 11th int'l conference passive and active measurement (PAM'10), (Berlin, Heidelberg). Springer, pp 41–50
12. Kaelbling LP, Littman ML, Cassandra AR (1998) Planning and acting in partially MServable stochastic domains. Artif Intell 101:99–134
13. Cai Y, Yu FR, Senarath G Optimal clustering and rate allocation for uplink coordinated multi-point (CoMP) systems with delayed channel state information(CSI). In: Proceedings IEEE ICC'13. Budapest, Hungary, June 2013

The ELF Tribe: Redefining Primary Education in a Post-COVID Era

G. Harish, B. Smitha Shekar, Anagha Ajoykumar, and K. Akshaya

Abstract Coronavirus brought the entire world to a grinding halt, and it has still not recovered from the vandalization caused. Along with the dramatic loss of lives worldwide, the situation also proved to be disruptive in every sector of the economy. Education system also faced immense turbulence. It has impacted the younger children the most as the effectiveness of online classes is highly doubtful among them. Even the teachers are not able to deliver classes to their best ability. Keeping in mind the present circumstances and the return to state of normalcy, 'The ELF Tribe' is designed as an educational tool for both teachers and students. There is substantial evidence in science to prove that kids learn better with hands-on experience but nothing can replace a good teacher. This application assists them in non-teaching tasks which, when blended with traditional schooling, has the potential to redefine meaningful schooling.

Keywords Online education · Teaching · Post-COVID · Python

1 Introduction

With the schools shut for more than a year due to the pandemic worldwide and having no other alternative to keep the academic years in progress, educators and students are forced to indulge to online learning. The most affected group are the primary school children, who are required to sit glued to the screens for hours. Quite ironically, the same parents and teachers, who once tried to keep the kids away from gadgets, are now insisting them to use the same as the indefinite duration of the outbreak gave no certainty about reopening of offline schools, waiting for which

G. Harish · B. S. Shekar (✉) · A. Ajoykumar · K. Akshaya
Dr. Ambedkar Institute of Technology, Bangalore, India
e-mail: smithashekarb.cs@drait.edu.in

G. Harish
e-mail: c_harishg.cs@drait.edu.in

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_11

would stunt the cerebral growth of children, as cognitive development happens the fastest in the growing years. Online education is the need of the hour.

This is also the golden era to make optimum use of the facilities available to mold the young minds into better human beings. Children have the power to change the world. Early child development sets the foundation for life. The entire course of humanity can be changed if kids are raised right in their formative years. For the longest of time, the education system has been blamed for its outdated approach and lack of significant impact. As the country strives to reform the existing structure, inculcation of technology could be the most efficient technique to create a significant change in the system. With a blended approach to schooling, we can completely reimagine and create a new and effective education model.

While most institution have interests in continuing to use the online teaching facility alongside offline class post the pandemic, the thought of continuing online classes in its current form does not sit well with the students. With complaints all over the country about the inefficiency of this teaching methodology, most of them cannot wait to get to physical classrooms. However, drawing inspiration from online schooling, if a mechanism is devised such that students do the same amount of curriculum in lesser amount of time at a better pace, and then invest the remaining time in other important skills, then the pandemic might pave way to a revolution in the Indian education system.

2 Scope of the Work

The ELF Tribe is proposed as an educational tool, a resource for students, teachers or even the management. It can be used in schools or under guided supervision, if at home. ELF is the abbreviation for 'Enjoy, Learn, Flourish,' implying that in order to reap the complete benefits of education, one must be able to enjoy the process. Unless the process is exciting, there is no desire to a promising result. It is important that children like learning as it helps them to stay curious, open to ideas and think out of the box. Without the drive to learn, the ability of the child to grow a full-fledged thinking is hindered. This tool makes learning a joyous process.

The proposed system tests the understanding of students without letting them feel the pressure through games and exercises. Teachers are relieved from all redundant tasks like attendance, assessment evaluation, etc., by automating them in the application, thereby get more time to invest in individual growth of each child. School management also utilizes the application to maintain records of both teachers and students. This is intended as a teaching material at school, which ensures limited screen time and controlled use of the application by children under the supervision of a teacher. Along with diverse assessment options, teachers can also track student activity and trace where they are lacking. The system hence advocates personalized learning and teaching mechanism.

3 Literature Survey

3.1 Evolution of Online Education

Though the pandemic played a huge role in normalizing the idea of online education and proliferating its usage even in the most basic levels like schools, the evolution of online learning dates back to the sixties. However, Internet was not an option then. Students of the University of Illinois, USA, used computer terminals to form a network. In 1983, Atari's former president, Ron Gordon, made online courses for those with access to personal computers by launching the Electronic University Network [1]. In 1984, University of Toronto made its first-ever course which was completely online. Three years later, the University of Phoenix initiated a fully online collegiate institution. It offered both undergraduate and postgraduate degrees and became the first institute in the world to do so.

In India, the broadcasting spaces like All India Radio and Doordarshan were known for telecasting recorded educational programs which were not only for higher education but also for school children. Many institutions like UGC, IGNOU and NCERT used the services provided by these national broadcasters, but there was always the need for interactive sessions since the classes were in recorded form. In 1994, the teleconferencing facility at IGNOU gave scope for live interaction which was a booster for a large number of online courses in India. It was also recognized as an official education channel under the Gyandarshan platform which was then made available in the DTH as GD-interactive channel.

3.2 National Education Policy 2020

July 2020 marked the witnessed a landmark decision in the field of education in the country when the Union Cabinet approved and passed the New Education Policy. This aims at making education from preschool to higher secondary level universal. Replacing the 1986 policy, this policy focuses on revamping and modernizing the curriculum, making the board exams easier, reducing a huge redundant chunk in the syllabus and retain only the core essentials and thereby, emphasizing on experiential learning and critical thinking in school education [2]. Though this is not the ultimate reform in the educational front after the last revision in the policy 34 years ago, it is definitely a start to restructuring the model.

With major reforms in the K-12 sector, one of the foundational principles that will govern the education system, according to the new policy, is the need for recognizing, identifying, and stimulate and nurture the unique potentials of each student, by prompting teachers and parents to foster each student's holistic development in both scholastic and non-scholastic arena [3]. Though this alteration in the ideology of education sounds exciting, it is difficult to achieve the desired outcome with traditional schooling alone. There is a need to incorporate blended methodologies into school education.

Motivation. This period is the beginning of a revolution whose potential was hidden for the longest of time. Establishing a two-way video communication between one to many was a huge challenge just fifteen years ago. With the need of social distancing and isolation today, online education is the only way to keep the academics. E-learning became the only option among the students and other stakeholders of the system across the world particularly during the pandemic, to keep up the momentum of the academic activities [4]. This has proven to be a period of momentous changes in education and has invariably seen a paradigm shift in the way even rudimentary education is perceived and delivered.

Information, communication and technology will undoubtedly aid to replace conventional teaching methodologies with technology-driven tools, facilities and amenities [5]. Though the current approach to online classes is not yet satisfactory, the recent developments in this domain are one of the most impeccable assets of the modern era. If harnessed well, it has the potential to catapult the entire world into great prosperity and development.

4 Design and Implementation

4.1 Data Flow Diagram

As depicted in Fig. 1, the proposed system has three kinds of users and hence, three components each for a different user, namely student, teacher and admin. Student portal and admin portal are desktop applications that can be installed in the computing devices of schools and accessed as per the time allotted and their requirements, respectively. In situations like these, it can be accessed from home if school permits. Teacher portal is a web application that can be accessed anytime by teachers. All components are independent entities connected to the same server and database. Hence, changes made in one portal are reflected in the others.

Admin portal has a login page for registered members of the school management with credentials preloaded in the database. Admin can register new teachers and students into the system and modify details of the existing teachers and students. They can also perform other critical tasks like allotting section for the student according to the academic year, allotting subject teachers' unique class and sections, or training the images required for students to login using face recognition. They also maintain the attendance of students with respect to the platform only.

Student portal has both manual login and using face recognition for the registered students. Manual login is used in situations where the face is not recognized due to some technical difficulty. Upon successful login, a drowsiness detection begins to run in the background, which rings an alarm if it detects droopiness. A voice-based

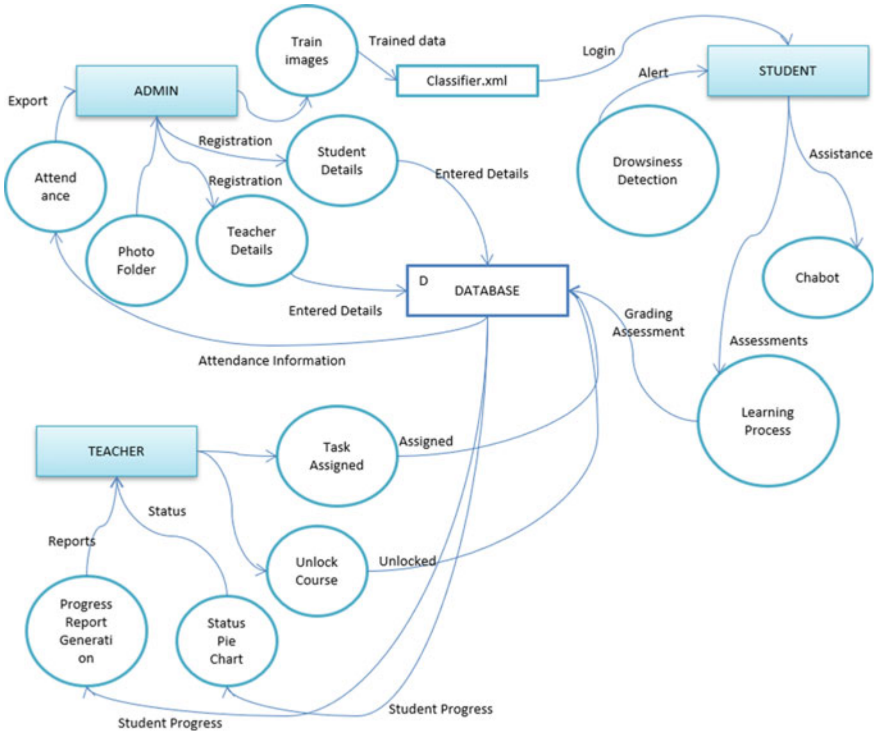
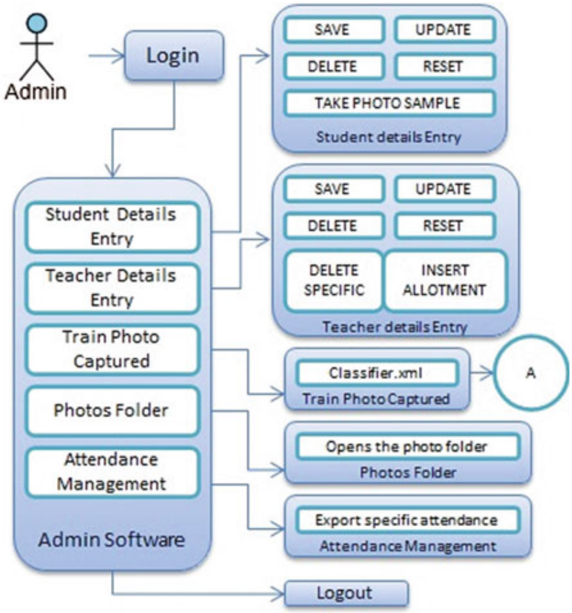


Fig. 1 Data flow diagram

virtual assistant helps them with queries in middle of a session. The daily tasks assigned by the teachers are displayed alongside the entire set of exercises related to their respective grade. It is presented in levels such that to attempt an exercise, all the previous ones in the hierarchy must be completed. This ensures that each student can proceed in their own pace. After every attempt, the scores are updated in the concerned teacher's view. If a student does not clear a level after three attempts, it gets locked until the teacher unlocks it for them. This tells teacher where the child needs help and hence ensures individual attention.

Teacher's portal has login facility only for teachers registered by the admin. They view the progress of the students they teach, assign daily tasks for them, unlock their locked courses and generate reports for each student. Class teacher can see information of all students of that class. A subject teacher can see only details of that subject. They can also see the aggregate performance of students for each module. This helps to analyses how well the topic is understood by the kids and helps them to focus more on those areas.

Fig. 2 Admin module



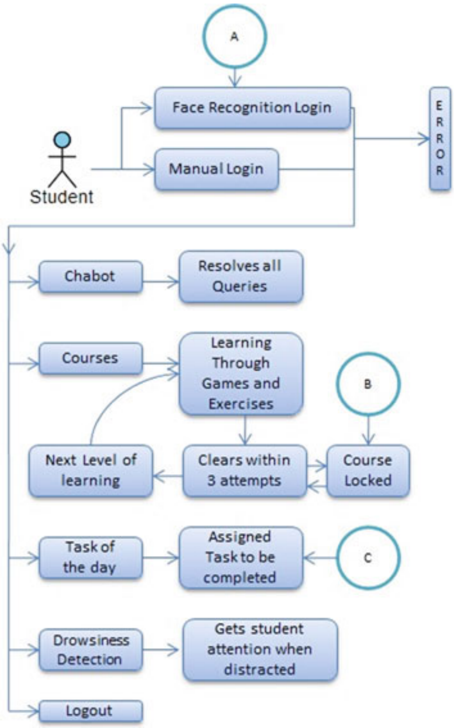
4.2 Schematic Diagram

Figures 2, 3 and 4 show the detailed schematic diagram of each user of the proposed system, i.e., admin, student and teacher.

4.3 Programming Technologies

The admin and student desktop applications are developed in Python, using Tkinter. The drowsiness detection system in the student module uses OpenCV. Haar Cascade is used for face detection [6]. The local binary pattern histogram (LBPH) algorithm is used for face identification [7]. The captured images are trained using LBPH algorithm to produce a classifier.xml file in the admin portal using which Haar Cascade algorithm detects the face in the student portal. The teacher's website is built using HTML, CSS and JavaScript. Google Charts is used to add a pie chart in the website to show the understanding of students in each module. XAMPP is used to mock the school server. The database is created using phpMyAdmin. To connect this database to the desktop applications, MySQL Connector package is used.

Fig. 3 Student module



5 Results and Discussions

5.1 Snapshots

5.2 Working

Admin creates records for teachers and student and allots classes and subjects to them. As a result, teacher views only students belonging to allotted class. It trains the data set for student face recognition by clicking real-time images of students in their physical presence. If the same face is tried to be registered in two student profiles, an error is shown that the image already exists. Admin has to access the physical location to delete the saved data set of a student. This helps in tracking the attendance of the students.

Teachers assign daily tasks to students based on what is taught in classroom. They can spot students or topics that need extra attention based on the scores of the exercises attempted. They generate reports for students. Entire history of student activity is available to teachers so as to keep track of what each student is doing in

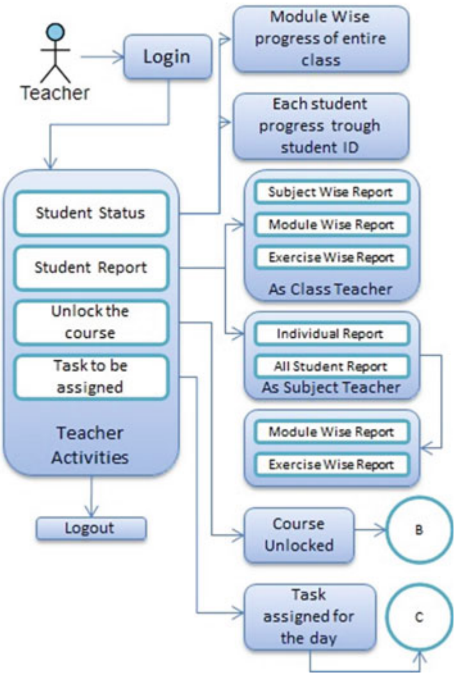


Fig. 4 Teacher module

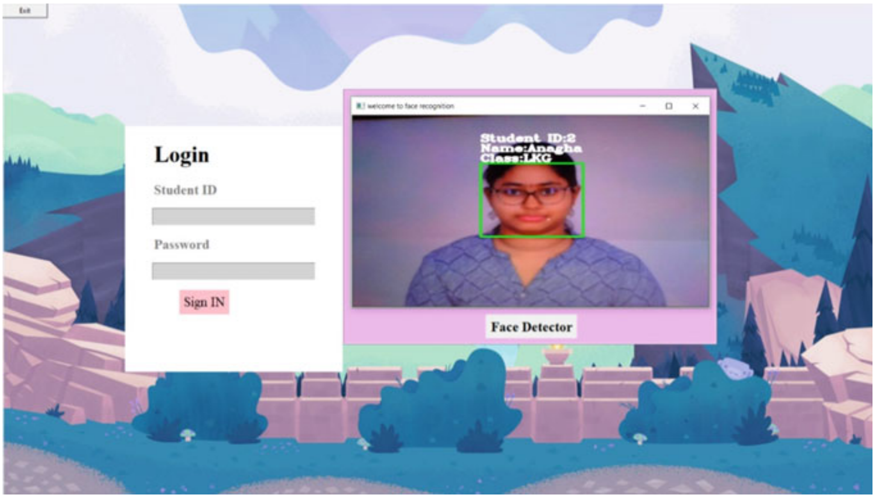


Fig. 5 Student portal login. Students can preferably login with face recognition as it is difficult for small children to type in their credentials. Manual login is incorporated as a backup in case the child needs help from the teacher

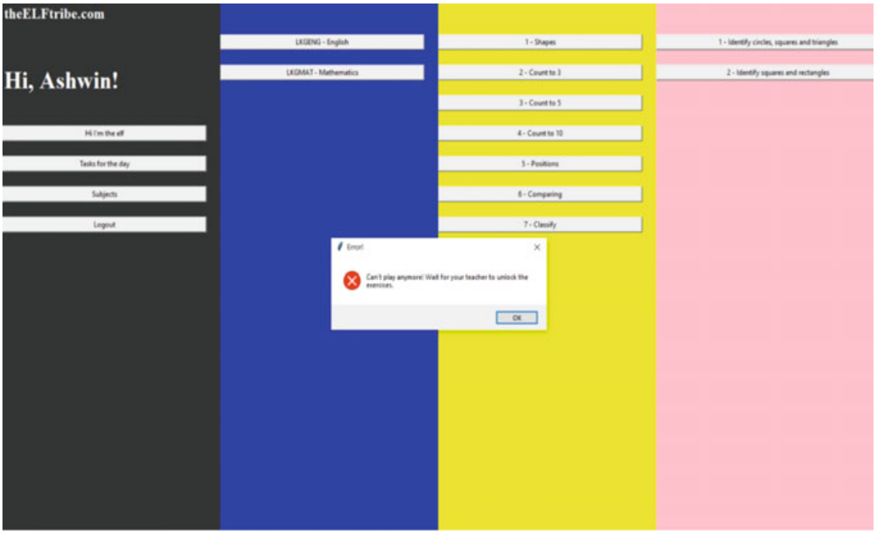


Fig. 6 Student have three attempts for each exercise. It gets locked after three failed attempts as there is no point in retrying until they get a better understanding of the topic. The teacher is notified about the same and can be unlocked by her only

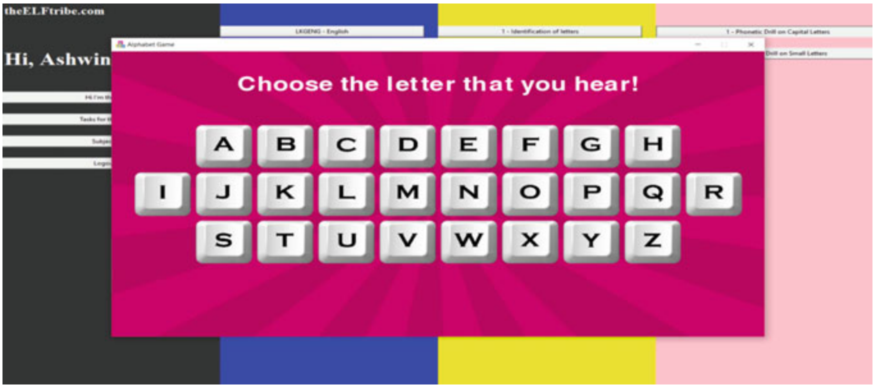


Fig. 7 English exercise for class LKG. It is a phonetic drill where the student must identify the letter that they hear and click the right one on the window. This tests both visual and auditory identification of alphabet

their portal. The mechanism provided for analyzing how effective their classes have been aids introspection and helps to improve the classroom experience for students.

Student portal has exercises for the respective academic year, structured in such a manner that the topic of an exercise requires complete understanding of the topic of the previous one and hence, students cannot proceed to next level until they finish all the previous levels. Virtual assistant clears doubts of students without the need to

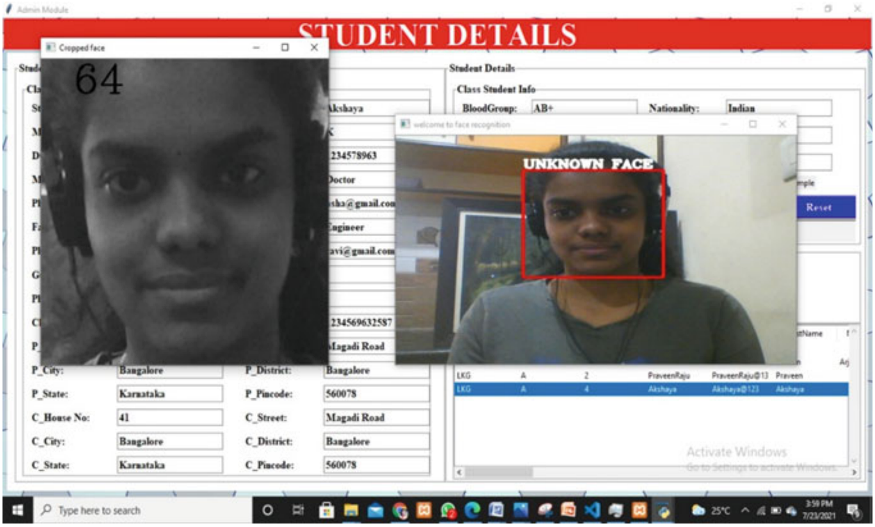


Fig. 8 Student registration and training images in admin portal. After obtaining information of the student is obtained during registration process, the student sits for capturing images and admin trains the data set for face recognition

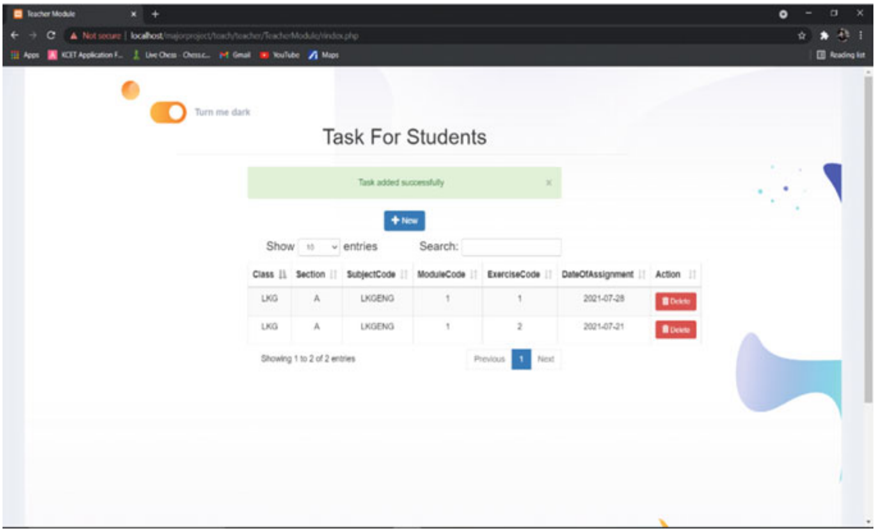


Fig. 9 Assigning task for student in teacher portal. Teachers can give daily tasks to the student based on what has been taught in classroom. They can assign or delete tasks only for students of the class and subject allotted to them by the admin

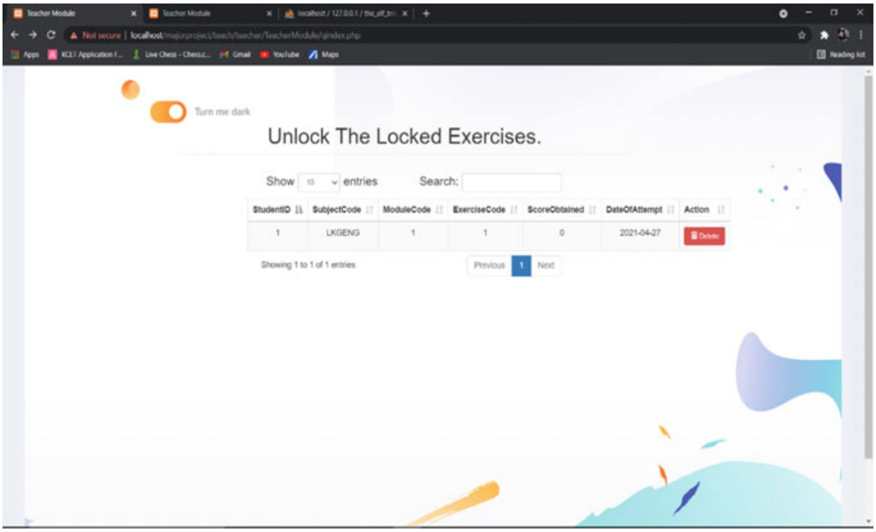


Fig. 10 Unlocking the locked exercises in teacher's portal. On being notified about the locked exercises of the students, teachers can unlock the locked exercises after making note of where they need help and student gets three fresh attempts

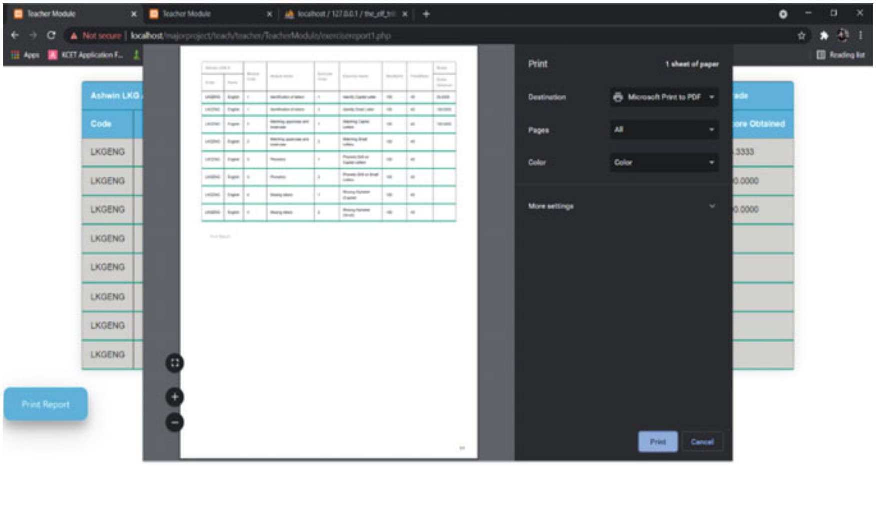


Fig. 11 Report generation in teacher's portal. Class teachers and subject teachers can generate reports of students either subject-wise, module-wise or even exercise-wise which can be either individual or for entire class

Fig. 12 Average performance of students in teacher's portal. The pie chart shows cumulative score obtained by students in each module, as average of all exercises in the module for self-analysis of the teacher to check how effective their classes have been

type, using audio-based bot. The drowsiness detection rings an alarm when the child is seen to be closing his eyes for a long time as feeling drowsy can lead to lack of concentration.

With the three components working in coordination, an effective resource is at the disposal of teachers, students and the school management to automate all mundane tasks and give them quality education where teacher is only required to teach and student to learn. There is no pressure to write or evaluate exams. Assessing happens with the ease of playing games, fastening the learning and provides time for other necessary activities like physical activity, social skills, spiritual learning and cocurricular activities, which are essential for facilitating the all-round development in a child. The work done here tries to achieve the same by assisting in imparting quality education.

6 Conclusion

The work is currently in its most primitive stage. To improve performance and expand its application to a wider audience, more creative modules can be added to enhance the enthusiasm of kids. Courses can be augmented for higher classes so as to implement this tool throughout schooling. More features can be incorporated into the coursework like emotion detection, image classification, etc. Lessons based on augmented reality, virtual reality or audio/video-based assessments can be included. In addition to the curriculum, moral sciences, environmental awareness, skills like social, cognitive, cultural, emotional, spiritual education or techniques to strengthen their consciousness can be included as these contribute to the overall development of the child.

Worker unions once opposed technological changes and machineries in factories as they saw no other way of living if their jobs got automated. In hindsight, those were the best inventions and life without them would have been impossible today. The scenario is similar in education today. Many argue that online learning or artificial intelligence will ruin the innocent minds of kids but in reality, AI can inculcate consciousness and intelligence in a child. They must do the academic part of the curriculum but the problem arises when it is the only thing for the twelve years of schooling. Technological catalysts like an AI-based educational tool aid all arenas of growth in kids.

The role of teacher has been slightly misinterpreted over the years. The value of a teacher in a student's life was immense in the olden days. The reason for this connection between them was the individual attention they gave each other. Teachers observed every child and guided them in all ways, be it academic or personal. They still have the power to make their future, but they are so burdened with other works that they do not have enough time to focus on every child. In time, it became so mechanical that there is absolutely no bond between them. Such tools are boons in such instances as well.

Our education system still has the remains of the colonial education system, which does not teach morals of life or humanistic values. They should be taught that money is neither a measure of success or happiness in life and that real achievements are in proving oneself as a person of values and ethics to the world. [8] Today's children are tomorrow's adults. The future of the country lies in the quality of our education system, and it should facilitate them to be better human beings. This work aims to be a helping hand in imparting the best education and assist them in this journey.

Acknowledgements Our sincere acknowledgment and a special thanks to the Management P.V.P Welfare Trust of Dr. A.I.T, for providing a congenial environment to work, learn and implement new technologies. Our sincere thanks to Dr. Siddaraju, Professor and Head, Department of C.S.E, Dr. A.I.T, for his support and continuous encouragement toward Education and Research and Development.

References

1. History of Online Education. https://thebestschools.org/magazine/online-education-history/
2. Aithal PS, Aithal S (2020) Analysis of the indian national education policy 2020 towards achieving its objectives. Int J Manage Technol Soc Sci (IJMTS) 5(2):19–41
3. Ministry of Human Resource Development, Govt. of India. https://www.education.go.in/sites/upload_files/mhrd/files/NEP_Final_English_0.pdf
4. Sathishkumar V, Radha R, Saravanakumar A, Mahalakshmi K (2020) E-learning during lockdown of covid-19 pandemic: a global perspective. Int J Control Autom 13:1088–1099
5. Ghavifekr S, Rosdy WAW (2015) Teaching and learning with technology: effectiveness of ICT integration in schools. Int J Res Educ Sci (IJRES) 1(2):175–191
6. Chinimilli BT, Anjali T, Kotturi A, Kaipu VR, Mandapati JV (2020) Face recognition based attendance system using haar cascade and local binary pattern histogram algorithm. In: 2020 4th international conference on trends in electronics and informatics (ICOEI)(48184), pp 701–704

7. Zhao X, Wei C (2017) A real-time face recognition system based on the improved LBPH algorithm. In: 2017 IEEE 2nd international conference on signal and image processing (ICSIP), pp 72–76
8. Kaushik A, Shah P, Chavan M, Dyer C, Ramachandran V, Sharma R (2009) Primary education in India: current status and future challenges. Vikalpa 34:61–90

An Extreme Machine Learning Model for Evaluating Landslide Hazard Zonation in Nilgiris District, Causative Factors and Risk Assessment Using Earth Observation Techniques

G. Bhargavi and J. Arunnehru

Abstract In Nilgiris mountain regions, landslide is one of the most severe hazards, damages natural resources, and disturbs the ecosystem and environment. It creates a substantial loss to agriculture, forest, wealth and scratches major architectural constructions in the mountainous region. This proposed work aims to distinguish the range of the spatial vulnerability of landslide using ensembled machine learning models, including AdaBoost, random forest algorithms, and gradient boosting decision tree algorithms. Seven conditioning factors namely slope angle, slope aspect, land use, geomorphic features, distance from road, drainage density, and lineament density and a total of 272 landslide historical locations with a ratio of 70/30 were used to construct the spatial database. Estimate the efficacy of the ensembled models using evaluation metrics like precision, recall, F1-score, MAE, MCE, MSE, Kappa and AUC score. Results concluded that the land-use factor, vegetable crop area, is prone to high vulnerability zone in the study area. Additionally, results show that among all the ensembled machine learning models, the AdaBoost algorithm's classification gives the highest accuracy value of 88.96%. However, the gradient boost decision classifier also outperformed with an accuracy value of 78.43%, and the random forest algorithm gives 72.61%.

Keywords Landslide susceptibility mapping · Machine learning · AdaBoost algorithms · Receiver operating characteristics · Nilgiris

1 Introduction

Landslides are uncertainties that create a loss of life, destruction to residences, highways, and road connections in mountain zones. Urbanized regions and other advancing projects endure more as the slopes go steeper to decline slope stability. The Nilgiris district hills are one of the familiar and widespread hilly regions in

G. Bhargavi (✉) · J. Arunnehru
Department of Computer Science and Engineering, SRM Institute of Science and Technology, Vadapalani, Chennai, Tamilnadu, India
e-mail: bg1064@srmist.edu.in

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_12

the southern part of India. The flow of time has experienced enormous growth and expansion with vast stretches of an area with tree plantations, coffee plantations, particular vegetable croplands, horticultural farms, and meadow lands for planting herb yields. In the Nilgiris district's study area, prolonged rainfall during monsoons and other human-made activities is the triggering factor for the landslides. As the report says there are 150 potential landslide locations distinguished after the landslide [1]. The preponderance of the landslides has happened in the tea plantation estates and other areas where crop's plantation was grown; from 1978 to 2021, many landslides occurred due to heavy rainfall in the region. Many families were buried, and the death rate was increasing in villages around the landslide region. Private transports were shattering away downhill sheer slopes. Geographic information system (GIS) is a powerful tool to identify landslide sensitivity zones. There exist several types of research implemented together with machine learning techniques, fuzzy logic and ANN techniques, logistic regression methods [2], the weight of evidence techniques [3], probabilistic representations, and frequency ratio classification [4]. This proposed work will use the most straightforward and effective method to produce landslide potential zone mapping, called the frequency ratio method, and machine learning classification algorithms.

2 Study Area

The Nilgiris district is a district in the southern part of India of Tamilnadu. Nilgiris is a part of the Western Ghats and one of the most magnificent mountain series pinpoint at the tri-junction border of Tamilnadu, Kerala, and Karnataka states of South India. Figure 1 shows the study area. The district, as mentioned above, has extended records of landslides [5]. The hazard occurs especially in October and November due to seasonal monsoon rain. The study area receives southwest monsoon and northeast monsoon rain, with 65% of them falling in railroad slip zones [6, 7].The literature

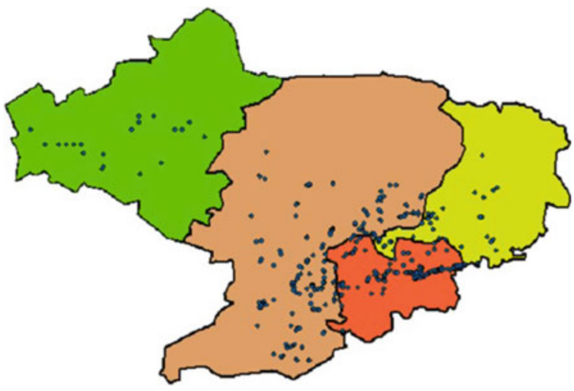


Fig. 1 Landslide points highlighted in blue dots

study reveals that most maximum of the landslides in the Nilgiris hill region has been triggered by extreme rainfall and anthropological and developmental projects.

3 Dataset

The collection of landslide analysis dataset is multiple approaches. [8]. We aimed to incorporate adequate data to interpret landslide susceptible zones in the Nilgiris district. The dataset collection is from various resources [9]. The study area persists below the cloud throughout the monsoon season. Hence, we need to collect cloud-free satellite data for our analysis [10]. But, these satellites fail to give precise data for the study area, which is under consideration. This proposed work considers various factors influenced for landslide includes slope angle, slope aspect, land use, geomorphic features, distance from road, drainage density, and lineament density. The dataset was taken from Sentinel-2 satellite images using the USGS website. Visual interpretation of this collected data for landslide mapping can accomplish using ArcGIS software. Finding out the position of earlier landslides and their nature are necessary for landslide susceptivity analysis [11, 12]. Yet, in the proposed work, the slides are interpreted as point data. The land- slide map prepared for the study area shows that the landslides should happen close to the highways and elevation ranging from 1800 to 2225 m.

4 Landslide Influencing Factors

The relationship between the landslide points and the thematic map layers is quantified as a frequency ratio [13, 14]. The spatial data of the determinant parameter, namely slope angle, slope aspect, land use, geomorphic features, distance from road, drainage density, and lineament density, are first using the appropriate measure to determine frequency ratio. The frequency ratios estimated for the several triggering factors are shown in Table 1.

4.1 Slope Degree

The slope represents a vital part in directing a region's resistance, and as the slope develops, the possibility of slope collapse rises [15, 16]. The surface wetness is high in such inclinations, and the closeness of river banks may cause landslides [17, 18]. Our study area classified the slope range from 0 to 60 ° into five different classes. In the Nilgiris district, there are more landslides potential points in 30 to 45 °.

Table 1 Landslide influencing factor for Nilgiris

S. No	Slope degree	No. of LSD (NLi)	LSD %	Area (No. of pixels) (NCi)	Area %	FR (FRi)	Normalization (FRn)
1	0–15 °	7	5.15	11,636	4.59	1.12	0.5
2	15–30 °	45	33.09	33,640	13.27	2.49	0.9
3	30–45 °	72	52.94	53,652	21.16	2.50	0.9
4	45–60 °	12	8.82	60,384	23.81	0.37	0.2
5	> 60 °	0	0.00	94,272	37.18	0.00	0.1
S. No	Slope aspect	No. of LSD (NLi)	LSD %	Area (No. of pixels) (NCi)	Area %	FR (FRi)	Normalization (FRn)
1	Flat	0	0.00	44	0.02	0.00	0.1
2	North	20	14.71	30,576	12.16	1.21	0.6
3	Northeast	12	8.82	28,672	11.40	0.77	0.4
4	East	24	17.65	33,548	13.34	1.32	0.7
5	South East	8	5.88	33,844	13.46	0.44	0.3
6	South	12	8.82	30,412	12.10	0.73	0.4
7	South West	16	11.76	29,004	11.54	1.02	0.5
8	West	32	23.53	32,416	12.89	1.82	0.9
9	North West	12	8.82	32,892	13.08	0.67	0.4
S. No	Land use	No. of LSD (NLi)	LSD %	Area (No. of pixels) (NCi)	Area %	FR (FRi)	Normalization (FRn)
1	Forest plantation	0	0.00	35,348	13.94	0.00	0.1
2	Dense forest	5	3.68	13,336	5.26	0.70	0.3

(continued)

Table 1 (continued)

3	Forest	22	16.18	50,392	19.87	0.81	0.4
4	Shrubs	0	0.00	12,820	5.06	0.00	0.1
5	Out crops	0	0.00	1292	0.51	0.00	0.1
6	Vegetable crop	98	72.06	72,996	28.79	2.50	0.9
7	Tea estate	6	4.41	63,664	25.11	0.18	0.2
8	Settlements	5	3.68	3736	1.47	2.50	0.9
S. No	Geomorphology	No. of LSD (NLi)	LSD %	Area (No. of pixels) (NCi)	Area %	FR (FRi)	Normalization (FRn)
1	Less dissected plateau	82	60.29	52,564	20.73	2.91	0.9
2	Moderately dissected plateau	28	20.59	143,132	56.44	0.36	0.1
3	Escarpment	9	6.62	7460	2.94	2.25	0.7
4	Valley fill	17	12.50	50,428	19.89	0.63	0.2
S. No	Distance to road (m)	No. of LSD (NLi)	LSD %	Area (No. of pixels) (NCi)	Area %	FR (FRi)	Normalization (FRn)
1	0–100	49	36.03	52,140	20.56	1.75	0.9
2	100–200	33	24.26	40,412	15.94	1.52	0.8
3	200–300	22	16.18	33,552	13.23	1.22	0.6

(continued)

Table 1 (continued)

4	300–400	14	10.29	26,600	10.49	0.98	0.5
5	400–500	5	3.68	21,604	8.52	0.43	0.2
6	Above 500	13	9.56	79,276	31.26	0.31	0.1
S. No	Drainage density (Km)	No. of LSD (NLi)	LSD %	Area (No. of pixels) (NCi)	Area %	FR (FRi)	Normalization (FRn)
1	0–2	8	5.88	31,316	12.35	47.63	0.4
2	2–4	36	26.47	47,640	18.79	1.41	0.1
3	4–6	40	29.41	50,508	19.92	147.67	0.9
4	6–8	28	20.59	85,816	33.84	0.61	0.1
5	Above 8	24	17.65	38,304	15.11	116.83	0.7
S. No	Lineament density (Km)	No. of LSD (NLi)	LSD %	Area (No. of pixels) (NCi)	Area %	FR (FRi)	Normalization (FRn)
1	0–0.5	59	43.38	61,128	24.11	1.80	0.9
2	0.5–1	28	20.59	58,572	23.10	0.89	0.5
3	1–1.5	39	28.68	101,300	39.95	0.72	0.4
4	1.5–2	10	7.35	23,196	9.15	0.80	0.5
5	>2	0	0.00	9388	3.70	0.00	0.1

4.2 Slope Aspect

The slope aspect distinguishes the declination track of the most significant rate of change of elevation in value from pixel to pixels. It exposes possible consequences of sweeping breezes, changeable climate conditions, and the occurrence of solar transmission [19]. According to the ground conditions, the adjustments that experience more intensive precipitation so that the soil becomes immersed more promptly, depending further on infiltration capability, managed by slope angle, vegetation and forest cover, and soil type [20].

4.3 Land Use

The influence of human activities affects the change in land use and land cover and modifications in the atmosphere [21]. We prepare the land use map by interpreting sentinel satellite data [22, 23]. In our study area, the plantation of vegetable crops considers a large area that covers more than 70% of the land use. The frequency ratio of this class is the highest value, and the normalized value is 0.9. At the same time, there are relatively insignificant slides in the settlement region, no slide points in forest regions.

4.4 Geomorphology

Diverse lineaments dissect the study area, and therefore, the contour regions can be segregating into common lineament fractures, moderate lineament fractures, and valleys (lands next to riverbeds and quickly submerged). In the proposed work, an escarpment is differentiated using a quadruple arrangement as a distinct unit. Tolerable dissected hill landform is the predominant class, accompanied by low lineament fracture slope, valley fill, and cliffs.

4.5 Distance to Road

The sloping edge for the road creates slope instability; 85% of landslides occurred within 400 m from the road. In particular, 35% of the landslides occurred within 100 m from the road distance. The study region within a range of 300 m has the highest normalization rate symbolizing a high possibility of landslides. Whereas, regions considerably away from the highway have a less normalized rate, which symbolizes a low possibility of landslide or non-landslide regions.

4.6 Drainage Density

Initially, the drainage canal leads to depletion of the sides and toe, provoking landslide. The maximum number of landslides has occurred within 100 m distance from streams, and therefore, normalization values within 100 m are always higher, indicating the high possibility of landslide occurrence. The high possibility zones lie between 4 to 6 km from the drainage distance in the proposed study area.

4.7 Lineament Density

Satellite data observes the lineaments are contours that delineate defects and fault zones (fracture in the terrain) [24]. Lineament is one of the primary triggering factors, especially in the Himalayan region, where two tectonic plates touch each other. For interpretation purposes, lineament split these density values into five different classes with a 50 m interval. The regions with a lineament density of 0 to 0.5 km are more sensitive to landslides. The rest of all the other classes are a low probability of landslides.

5 Landslide Susceptibility Modelling

Specifically, our model utilized three distinct ensemble machine learning techniques to classify the landslide dataset, namely AdaBoost, random forest, and gradient boosting decision tree [25]. The machine ran 70% of the training data and 30% of the test data with various parameter combinations using machine learning algorithms to train the model. Then, evaluate the model using various performance evaluation techniques and calculated the coefficient of variation to estimate the possibility of the consequences. The descriptions of the three machine learning techniques are described in detail. Figure 2 shows the methodology for the proposed work.

5.1 AdaBoost

Freund and Schapire introduced AdaBoost [26], one of the conventional boosting strategies, including significantly improved application and adaptive resampling procedure as it regulates bias and variance [27]. AdaBoost trains the many weak classifiers to train the specific dataset and then consolidate these weak classifiers to develop a robust classifier. Among the several weak classifiers, a low error rate classifier is picked as the final classifier.

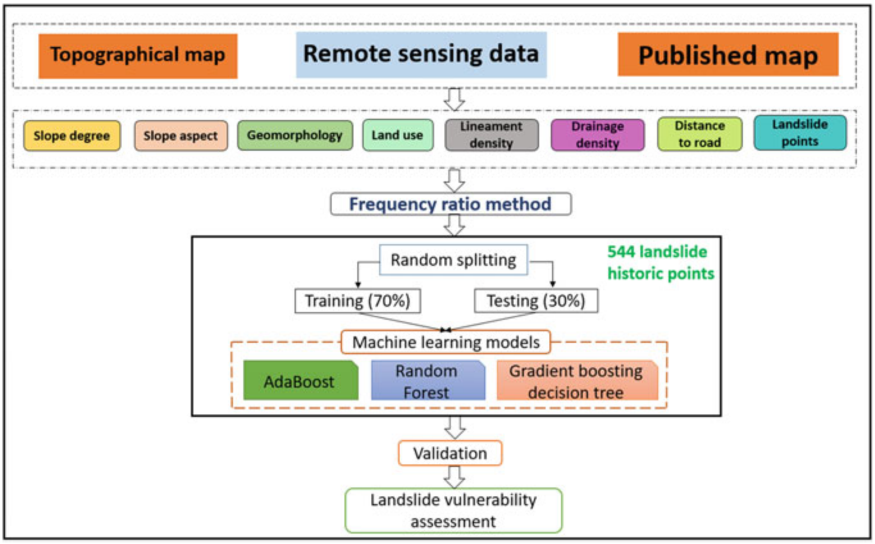


Fig. 2 Landslide susceptibility flow diagram

$$F(a) = sign(\sum_{1}^{n})C_m f_m(a) \tag{1}$$

The F(a) denotes the final classifier, the weighted sum of every *M* iteration, *C* value, and the classifier output in this equation. *M* indicates the weak classifier, fm(a) is the product of the weak classifier. *C* is the weight involved in the classifier as applied by AdaBoost. This method can overcome bias and variance, enhance classification ability, and possess more superior effectiveness. AdaBoost is receptive to eccentricity, so it affects the accuracy of the decisive classifier. In this proposed work, the iteration runs for 500 epochs where the training rate is 0.8.

5.2 *Random Forest Algorithm*

The random forest refers to ensemble learning techniques accumulating many classification and regression trees (CART) to classify and predict [28]. Random forest is one of the predominant ensemble algorithms in multi-classification and prediction analysis[37]. The advantage of a random forest classifier is, it does not require scaling, hyper-parameter tuning, and transforming input data.

The random forest contracts with dropping values, and also, it can continue eccentricity in predictive variables. It combines random subspace and bagging ensemble learning technique which is counters over-fitting problems.

5.3 Gradient Boosting Decision Tree

Gradient boosting is the same as AdaBoost; both boosting techniques use an ensemble of decision trees to forecast a target variable. However, the depth of the gradient boost trees is higher than 1.The primary objective of this gradient boosting is to generate diverse weak classifiers and subsequently consolidate them to build a robust classifier [29]. The landslide forecast outcomes conclude on the number of weak classifiers. It takes a set of 100 iterations and 1 data per learning rate.

6 Model Performance Analysis and Accuracy Evaluation

The proposed work, landslide risk susceptibility modelling and vulnerability zone mapping a classification problem, with binary results stating the occurrence of landslide and non-landslide locations. In this classification model, the outcome class is either represents the landslide occurrence or non-landslide zones. The evaluation metrics accuracy (ACC) is a spontaneous performance measure that shows the percentage value of the accurately predicted value to the total observations Eq (1).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{2}$$

We have got 0.8896 which implies our model is approx. 88% accurate. Precision measures the exactness, and it shows the percentage value that predicts the positive values from the total observation Eq. (2). This metric gives the number of landslide occurred from the total number of landslide point mentioned.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

The receiver operator characteristic (ROC) curve is one of the essential evaluation metrics. A probability curve plots the true positive rate (TPR) against the false positive rate (FPR) at varied threshold rates and typically classifies the correct predicted value from the total [28]. Accuracy describes the amount of accurately labelled samples of both landslide and non-landslide occurrence, varying from 0 to 1, and the greater the AUC values, the more reliable the model functions. Table 2 shows the model

Table 2 Various performance metrics evaluation for machine learning models

ML models	Precision	Recall	F1-score	MAE	MCE	MSE	Kappa	AUC Score
AdaBoost	0.88	0.87	0.87	0.13	0.74	0.13	0.73	0.95
RF	0.87	0.87	0.87	0.13	0.73	0.13	0.73	0.94
GBDT	0.86	0.86	0.85	0.14	0.71	0.14	0.70	0.92

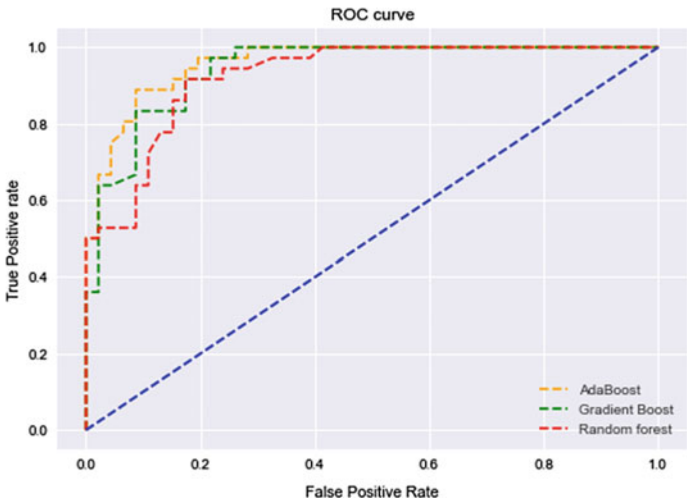


Fig. 3 ROC curve displays the performance of classification models

evaluation results with various performance metrics, indicates that three machine learning algorithms were performed equally well for landslide classification.

7 Model Validation and Comparison of Results

Analysing machine learning techniques and choosing an ultimate model is a standard process in applied machine learning. Construct the model using three machine learning techniques: AdaBoost, random forest, and gradient boosting decision tree algorithm. The AdaBoost algorithm achieves the highest accuracy value of 88.96% and compares the other metric values to justify the best accuracy model. The model is evaluated using precision, recall, and F1-score of the AdaBoost is 0.87. The AUC score value is 0.95, which is comparatively high than other models. Gradient boost decision classifier gives 78.43%, and random forest algorithm gives 72.61%. Figure 3 shows the ROC curve of the three machine learning algorithms.

8 Conclusion

The hazard is challenging to predict, whereas distinguishing the range of landslides in vulnerability zones using the historic landslide points will be the significant step for assessing the risk factors. Prepare a landslide susceptibility map using seven primary triggering factors. Classify the parameters into five sections using the natural breaks technique in the ArcGIS tool. Frequency ratio formula has been applied to develop the landslide susceptibility map for the Nilgiris district using historic landslide points. The geospatial dataset has randomly divided into a training (70%) and testing datasets (30%). If the landslide point is one, then the sample data is the landslide occurred zone; otherwise, the class value is "0". Build a machine learning model using three different ensembled machine learning algorithms: AdaBoost algorithm, gradient decision tree, and random forest algorithm. After evaluating the ensembled models using various metrics, the AdaBoost algorithm holds the highest accuracy value of 88.96%. In this study area, the vegetable crop is the high vulnerability zone with a 30–45 ° slope angle. This area has more lineament fracture plateau, which is the other major reason for hazard occurrence.

References

1. Saranaathan SE, Mani S, Ramesh V, Venkatesh SP (2021) Landslide susceptibility zonation mapping using bivariate statistical frequency ratio method and gis: a case study in part of Sh 37 Ghat road, Nadugani, Panthalur taluk, the Nilgiris. J Indian Soc Remote Sens 49(2):275–291
2. Bhargavi G, Vaijayanthi S, Arunnehru J, Reddy PRD (2021) A survey on recent deep learning architectures. In: Artificial intelligence and IoT: smart convergence for eco-friendly topography, pp 85
3. Dahal RK, Hasegawa S, Nonomura A, Yamanaka M, Dhakal S, Paudyal P (2008) Predictive modelling of rainfall-induced landslide hazard in the lesser Himalaya of Nepal based on weights-of-evidence. Geomorphology 102(3–4):496–510
4. Bhargavi G, Arunnehru J (2020) Land risk susceptibility, hazard and risk factors in western ghats, india–a review. J Crit Rev 7(19):10020–10028
5. Ganapathy GP, Rajawat AS (2015) Use of hazard and vulnerability maps for landslide planning scenarios: a case study of the Nilgiris India. Nat Hazards 77(1):305–316
6. Thennavan E, Ganapathy G, Chandrasekaran S, Rajawat A (2020) Probabilistic rainfall thresholds for shallow landslides initiation—a case study from the Nilgiris district, Western Ghats. India. Int J Disaster Risk Manage 2(1):1–14
7. Colangelo G, Perrone A (2012) Geoelectrical tomography as an operative tool for emergency management of landslide: an application in Basilicata region, Italy. Int J Geophys
8. Chandrasekaran SS, Elayaraja S, Renugadevi S (2009) Damages to transport facilities by rainfall induced landslides during November 2009 in Nilgiris, India. In: Landslide science and practice, pp 171–176. Springer
9. Jayanthi J, Raj TN, Gandhi MS (2016) Identification of landslide—prone areas using remote sensing techniques in Sillahallawatershed, Nilgiris district, Tamilnadu India. Int Res J Eng Technol 3(6):1947–1952
10. Pardeshi SD, Autade SE, Pardeshi SS (2013) Landslide hazard assessment: recent trends and techniques. SpringerPlus 2(1):1–11
11. Guzzetti F, Mondini AC, Cardinali M, Fiorucci F, Santangelo M, Chang KT (2012) Landslide inventory maps: new tools for an old problem. Earth-Sci Rev 112(1–2):42–66
12. Lv ZY, Shi W, Zhang X, Benediktsson JA (2018) Land-slide inventory mapping from bitemporal high-resolution remote sensing images using change detection and multiscale segmentation. IEEE J Sel Top Appl Earth Observations Remote Sens 11(5):1520–1532
13. Jennifer JJ, Saravanan S, Abijith D (2021) Application of frequency ratio and logistic regression model in the assessment of landslide susceptibility mapping for Nilgiris district, Tamilnadu, India. Indian Geotech J, pp 1–15
14. Wieczorek GF (1984) Preparing a detailed landslide-inventory map for hazard evaluation and reduction. Bull Assoc Eng Geol 21(3):337–342

15. Rahamana SA, Aruchamy S, Jegankumar R (2014) Geospatial approach on landslide hazard zonation mapping using multicriteria decision analysis: a study on Coonoor and Ooty, part of Kallar watershed, the Nilgiris, Tamilnadu. Int Arch Photogrammetry Remote Sens Spatial Inf Sci 40(8):1417
16. Guru B, Veerappan R, Sangma F, Bera S (2017) Comparison of probabilistic and expert-based models in landslide susceptibility zonation mapping in part of Nilgiri district, Tamil nadu, India. Spat Inf Res 25(6):757–768
17. Rajakumar P, Sanjeevi S, Jayaseelan S, Isakkipandian G, Edwin M, Balaji P, Ehanthalingam G (2007) Landslide susceptibility mapping in a hilly terrain using remote sensing and gis. J Indian Soc Remote Sens 35(1):31–42
18. Senthilkumar V, Chandrasekaran SS, Maji VB (2017) Overview of rainfall induced landslide events and importance of geotechnical investigations in Nilgiris district of Tamilnadu, India. In: 2017 workshop on world landslide forum, pp 281–287. Springer
19. Tassano M, Montanez A, Nunez L, Trasante T, Gonzalez J, Irigoyen J, Cabral P, Cabrera M (2021) Spatial cross-correlation between physicochemical and microbiological variables at superficial soil with different levels of degradation. CATENA 198:105000
20. Grossman RL, Durran DR (1984) Interaction of low-level flow with the Western Ghat mountains and offshore convection in the summer monsoon. Mon Weather Rev 112(4):652–672
21. Jaiswal P, van Westen CJ, Jetten V (2010) Quantitative landslide hazard assessment along a transportation corridor in southern India. Eng Geol 116(3–4):236–250
22. Venugopal D (2004) Development-conservation dilemma in the nilgiri mountains of south India. J Mt Sci 1(1):74–80
23. Neelakantan R, Yuvaraj S (2013) Relative effect-based landslide hazard zonation mapping in parts of Nilgiris, Tamilnadu, South India. Arab J Geosci 6(11):4207–4213
24. Youssef AM, Pourghasemi HR (2021) Landslide susceptibility mapping using machine learning algorithms and comparison of their performance at Abha basin, Asir region. Saudi Arabia Geosci Frontiers 12(2):639–655
25. Yu H, Lu Z (2018) Review on landslide susceptibility mapping using support vector machines. CATENA 165:520–529
26. Bui DT, Tsangaratos P, Ngo PY, Pham TD, Pham BT (2019) Flash flood susceptibility modeling using an optimized fuzzy rule based feature selection technique and tree based ensemble methods. Sci Total Environ 668:1038–1054
27. Pham BT, Bui DT, Prakash I, Dholakia MB (2017) Hybrid integration of multilayer perceptron neural networks and machine learning ensembles for landslide susceptibility assessment at Himalayan area (India) using gis. Catena 149:52–63
28. Arunnehru J, Geetha MK (2016) Difference intensity distance group pattern for recognizing actions in video using support vector machines. Pattern Recogn Image Anal 26(4):688–696
29. Arunnehru J, Geetha MK (2013) Behavior recognition in surveillance video using temporal features. In: 2013 fourth international conference on computing, communications and networking technologies (ICCCNT), pp 1–5. IEEE

Analysis of Cross-Site Scripting Vulnerabilities in Various Day-To-Day Web Applications

Awani Kendurkar, Jayesh Mandlekar, N. Jeyanthi, and R. Thandeeswaran

Abstract Networking threats aim to disrupt the normal flow of data and communication. They attack basic security measures—confidentiality, integrity and availability. Cross-site scripting (XSS) aims at rattling either confidentiality or integrity, depending on the focus of the attack. Since the dawn of the Internet, the amount of cyber-attacks and also the need for cybersecurity has grown exponentially. Today, there are numerous predefined ways network resources that can be attacked. It can be via malware or networking threats like XSS. This paper is an attempt to discover the most commonly exploited flaw in today's web pages—the XSS vulnerability—using various testing tools like Burp Suite and Nessus.

Keywords Cross-site scripting (XSS) · Confidentiality · Real-time applications · Threats · Attacks · Burp Suite · Nessus

1 Introduction

Networking threats aim to disrupt the normal flow of data and communication. Security Triad C-I-A is compromised, and data breached. XSS aims at rattling either confidentiality or integrity, depending on the focus of the attack. In the most basic implementation, XSS injects attacking code (malicious or not) into the poorly written HTML document of any web application. This is possible by including the < script > tags into an input field with appropriate knowledge of interior JavaScript code. This

A. Kendurkar · J. Mandlekar · N. Jeyanthi (✉) · R. Thandeeswaran
School of Information Technology & Engineering VIT, Vellore, Tamil Nadu, India
e-mail: njeyanthi@vit.ac.in

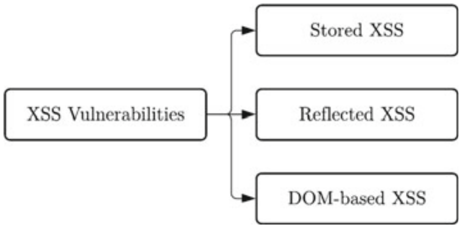
A. Kendurkar
e-mail: awani.kendurkar2018@vitstudent.ac.in

J. Mandlekar
e-mail: jayesh.mandlekar2018@vitstudent.ac.in

R. Thandeeswaran
e-mail: rthandeeswaran@vit.ac.in

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_13

Fig. 1 Types of XSS



is then loaded by the web application at the next instance of use and thus disturbs the natural progression of events.

Although it seems simple at first, writing XSS code certainly is a skill. But when executed, it can be of huge implications. It opens up avenues of grave consequences for websites providing financial, health care or other services that involve the flow of sensitive data. We aim at understanding this skill, practising it and applying knowledge to a real-world scenario (possibly a popular web application).

Bounties in the network security sphere are rewards offered to researchers that find pitfalls within developer codes and report it to the corporation—without leaking the existence of vulnerability first. The bounties depend on what the company can offer, from cash to privileges and even lead to being a viable job [1].

Coming back to XSS, it is essential to note that there is not much actual programming from the researcher's side, and it is more of an analysis job. The security researcher navigates through the application's source code and looks for bypasses that can be used as gateways to exploitations.

Following the discovery, malicious code (in the research aspect, test code) is injected via the bypass hence found and triggered to receive a response. This code is called a vector, which is designed keeping in mind the restrictions imposed by various aspects of a webpage (such as input sanitization, event handling, web application firewalls—WAFs, and even the browser itself). As such, there are three types of XSS vulnerabilities, as shown in Fig. 1 and explored in Table 1.

2 Literature Review

Engin et al. [2] point out that HTML encoding can be of variety—and prevention and detection of XSS at the time was (or even today is) extremely difficult, the method does not defend against XSS-based covert attacks. Vogt et al. [3] acknowledge that most XSS defence techniques were based on server side, and applications are not fixed until several users are infected via client-side attacks. This method focuses more on securing private/sensitive information—by tracking the flow. Bisht and Venkatakrishnan [4] identify that client-side input filtering and validation are not enough to prevent attacks. It mentions "Samy" computer worm—spread on MySpace and shut it down within 20 h. It basically wrote a script within the site that added "Samy" as a friend and replicated itself onto the visitor's profile. The defence method

Table 1 Types of XSS

XSS type	What it does	Where it is	Also called
Stored XSS	When the server requests the stored data, the malicious script is retrieved from the server	Target servers store the malicious script indefinitely	Persistent or Type-I XSS
Reflected-XSS	The victim's browser executes the code reflected back since it came from a "trusted" source	The web server reflects the code	Non-persistent or Type-II XSS
DOM-based XSS	A change in the victim's browser's DOM "environment" results in the execution of the attack payload	Client-side code executes abnormally as a result of the injected code; HTTP response remains unchanged	Type-0 XSS

identifies unauthorized scripts and prevents output on the response on the server side. The crux of the framework generates and hosts a shadow page of the application and checks for any changes to the actual page. If there are any flaws, the request is denied from being succeeded.

Michael and Monica [5] propose a goal-directed model-checking system that generates attacks and deploys onto web applications in a safe environment. Analysts can specify the target and method of attacks to test the system reports back threats that were successful in exploitation. The testing provided in the research is not adequate by today's standards (mere 130,000 lines of codes). Sun et al. [6] propose a purely client-side solution to the mitigation of XSS attacks by identifying replication attacks. The method monitors requests and checks for ones that send self-replicating modules, made possible by checking the embedded scripts. It had been deployed as a Firefox extension and claims low-performance overhead without false positives in their testing.

Weinberger et al. [7] conduct an analysis on current security frameworks that work on XSS sanitization—developing a new model of the web browser for testing and characterizing select challenges of XSS sanitization. This analysis presented major flaws exploitable by XSS in 14 major frameworks and large web applications. Machine learning has been used for the prevention and detection of other cybersecurity vulnerabilities and how they are simple and straightforward by fabricating an algorithm that extracts basic and contextual features of a web application source code and builds a model to predict any threats [8]. Machine learning algorithms solve the problem of XSS vulnerabilities. The solution in [9] focuses on social-networking sites and points out that the current prevention methods are not adequate due to the amount of JavaScript and AJAX included in these applications.

According to [10], filtering proxy server intended to be placed at the web application's site in order to filter both incoming and outgoing data streams has several limitations: 1. Simple use of regular expressions or malicious encoding in hexadecimal format (or more advanced) can be used to avoid filters placed as a defence by skilled attackers. 2. Proposed semantics of the policy language is error-prone because it is not duly reported. 3. The filtering proxy server can result in performance and scalability issues during deployment. 4. Only addresses JS-based XSS attacks even though several other languages like Java or Flash should be considered.

Various significant concerns for web applications of diverse organizations like banking, health care, financial services were analysed by referring Website Security Statistics Report of WhiteHat Security. Gupta and Gupta [11] claim that 55% of banking websites are "always vulnerable" to cyber-attacks. Two constraints for the current state of website security: 1. Frequency of modification of source code of web applications. 2. Static analysis of source code of web applications.

An attribute clustering algorithm is used, which is based on the k-means clustering algorithm in [12]. The time complexity is calculated as *O(n x m^3logm),* where *n* and *m* are dimensions of the original dataset. Low rates of false positives and false negatives were found [13] during the HTML5 detection phase of the attack vectors on the web servers incorporated into the implicit host systems of the ICAN cloud simulator. The statistics of the exploited attack vector on various HTML5 web applications like WordPress, Drupal and Joomla reveal 128 true positives out of 147 attacks for WordPress, more than any other web application tested. The vulnerability is prevented by exploiting the automated placement of sanitizers in the injected JS code [14]. The motivation arose when it was observed by the authors that input validation is the first defensive mechanism that is deployed on the server side to mitigate XSS attacks, even though it is ineffective when an attacker injects feature-rich HTML content.

This detection and prevention system proposed is based on encoding the HTML/JS keywords by randomization techniques which distinguish regular script from the malicious script injected [15]. Statistics to remedy cookie thefts are presented in [16] in detail. Features are extracted using the Word2vec NLP technique in [17]. Separate Gaussian mixture models are utilized to characterize XSS and routine web transactions, respectively, and integrate predictions from these models to provide a better classification of a new web transaction. A prototype developed in Java and installed as an extension on Google Chrome [18] based on a JavaScript string comparison and context-aware sanitization detects the injection of XSS worms, also determines the context of these worms and performs sanitization accordingly.

3 Prevention and Mitigation Methods

A detailed study of all the proposed methods reveals mainly two kinds of mitigation methods for XSS vulnerabilities. Although traditional security techniques include firewalls, cryptography-based mechanisms, secure coding/programming practices and input validation, there are two broad methods that are used widely by web applications to avoid being susceptible.

1. *Analysis and filtering of the data exchanged*: A filtering proxy is intended to be placed at the server-side of a web application to differentiate trusted and non-trusted traffic into separated channels. This can also filter both incoming and outgoing data streams. Malicious URLs can also be distinguished on the client side by blacklisting them and preventing the victims' browsers from contacting them. The client-side proxy then rejects the redirection to these URLs. However, this method is not likely to detect complex XSS attacks. The use of proxies, in general, is easy in theory and proposal but complicated in practice. The deployment is complex, especially for applications with high client-side processing. The performance and scalability of such web applications might also be affected adversely due to this practice.

2. *Runtime enforcement of web browsers*: An auditing system for the web browser Mozilla Firefox's JS interpreter can be developed, which is based on an IDS system that detects misuses during the execution of JS operations. Countermeasures can be taken to avoid violations against the browser's security if there is an abuse of browser resources, like the transfer of session cookies associated with the web application's site, to untrusted parties. Since the application of these countermeasures can introduce scalability issues, an alternative is to issue a simple warning message to the user that the resources are being transferred to a third party that is not necessarily trustworthy. This method adds a layer of security but under the final decision of the user, who might not be aware of the risks involved in such a transfer.

An essential technique of mitigating XSS attacks remains carrying out safe input handling. Encoding mechanisms should be exploited each time user input is incorporated into a web page. In several other cases, encoding mechanisms have to be substituted with input validation mechanisms. Safe handling of user-supplied input has to take into consideration which perspective of a web page the user-supplied input is injected into. Therefore, to thwart all types of XSS attacks, safe handling of user-supplied input has to be exploited in both client-side and server-side source code of web applications.

4 Analysis of XSS Attacks

Table 2 provides a concise view illustrating recent XSS attacks on popular web applications like Facebook, Twitter and LinkedIn.

5 Implementation

5.1 Burp Suite

Burp Suite provides a range of cybersecurity tools [19], which brought to the users by PortSwigger. PortSwigger is a research-driven community that offers tools for

Table 2 Analysis of XSS attacks on popular web applications

Target	Methodology	Mitigation
Facebook (2020)	Exploits window.open ('java script:alert(document.domain)') in the code, present as a DOM vulnerability	Facebook fixed this by adding facebook.com regex domain and schema check in the payload URL parameter
CIA (2011)	DOM-based vulnerability leading to the defacing of the website	Implementation of basic threat detection should have been in place before such an attack
Yahoo (2014)	A DOM-based vulnerability that led to the hacking of Yahoo accounts via infected links or receipt of spam	Requires input sanitization and detection methods for avoiding such attacks
Twitter (2014)	Attacked the HTML parser of the Twitter tweeting textbox, which allowed the attacker to write additional JavaScript code which was unfiltered—giving rise to the creation of a worm	HTML parsers, once fixed, can prevent such kinds of attacks in future
PayPal (2018)	URL encoding/decoding onto PayPal's URL gives access to XSS vulnerability which can be exploited remotely	PayPal responded to the bounty hunter by fixing the vulnerability, rewarding the security researcher, and producing a report
UK parliament website (2014)	Flaws discovered in the website's search engine—allowing exploiters to get access to confidential media, as well as passwords	The XSS flaw was disclosed to the UK Parliament, and the government confirmed that a fix had been put in place
Hotmail (2011)	A vulnerability that allows the attacker to steal keystrokes, cookies, or other pieces of sensitive information. Simply hovering on the email preview led to account compromise	Mishandled HTML strings enabled attackers to gain unauthorized access to information, or the ability to take actions while posing as the user
Google groups (2020)	Stored XSS is activated when a set of attack files are uploaded to a group. Users that access these files are then vulnerable to getting their cookies (sensitive information) leaked	Input sanitization would have to take place at a file-level for such an attack
LinkedIn (2018)	A common vulnerability that was taken advantage of—by using the Embed exploit technique for the "articles" part of the website. An attacker is able to recreate and embed an innocent-looking Linked login prompt or other phishing strategies within articles, which gives the attacker access to sensitive information	Embedding technique should have some kind of verification for any third-party access from scripts within such features

(continued)

Table 2 (continued)

Target	Methodology	Mitigation
Amazon Alexa (2020)	The gap in the system gives attackers access to voice history, personal and bank information. This is due to the misconfiguration of the CORS policy, which allowed AJAX requests to be sent from Amazon subdomains	A unique instance of XSS attack to mitigate since voice-control technology is relatively new. A way to stop such an event from happening is to secure "Cross-Origin Resource Sharing"

web application security, testing and scanning. Burp Suite is an ensemble of tools used globally and popularly for penetration testing, bounty hunting, research as well as a hobby. Within our project, we look to use Burp Suite as a tool to identify possible vulnerabilities and routes within our target application, but due to budgeting restrictions, we are unable to unlock all required features to utilize this suite completely.

5.2 Vulnerability Scanning via Nessus and Burp Suite

Before proceeding to actual testing of attack vectors on the websites, we try to list the identifiable vulnerabilities using a scanner such as Nessus and Burp Suite, which tracks the network activity. A list of potential targets is JioMart (www.jiomart.com), BigBasket (www.bigbasket.com), PayTM (www.paytm.com) and VTOP (www.vtop.vit.ac.in).

The above websites were chosen randomly, but also with a sense of concern regarding their securities. JioMart and BigBasket are the largest supermarkets with online delivery options; PayTM is the most popular wallet and UPI application in India, and VTOP is the student login portal of our college (Vellore Institute of Technology). Any and all security vulnerabilities have our personal interest due to the regularity with which we interact with them on a daily basis.

The Nessus scan results mainly were unsatisfactory. The HTTP issues found in VTOP and BigBasket included HSTS Missing From HTTPS Server, HTTP Methods Allowed. No traces of detectable XSS vulnerabilities were found on the JioMart website, and PayTM did not yield any vulnerabilities.

Burp Suite itself provides a large set of tools for vulnerability testing. However, due to budget restrictions, we are constricted to using the community edition, which can only listen to HTTP requests/responses and proxies. No such XSS vulnerability was identified from this analysis. However, there is one critical vulnerability to be noted that is not relevant to the discussion about XSS. Going through HTTP requests/responses of VTOP, it was found that the username and password of a user

```
POST /vtop/doLogin HTTP/1.1
Host: vtop.vit.ac.in
User-Agent: Mozilla/5.0 (X11; Ubuntu; Linux x86_64; rv:81.0) Gecko/20100101 Firefox/81.0
Accept: */*
Accept-Language: en-US,en;q=0.5
Accept-Encoding: gzip, deflate
Content-Type: application/x-www-form-urlencoded; charset=UTF-8
X-Requested-With: XMLHttpRequest
Content-Length: 524
Origin: https://vtop.vit.ac.in
Connection: close
Referer: https://vtop.vit.ac.in/vtop/initialProcess
Cookie: JSESSIONID=0751AA0250C8BBB6500C3E3CC1560236; loginUserType=vtopuser; SERVERID=s1

uname=18BIT0044&passwd=REL[redacted]r&g-recaptcha-response=
03AGdBq25IqOhHrvK_84bx5uhPxGE0MwoDB4rZNq94r6-yrrJ800vfPporHoFDzYO7XoFIijoT2NN2vSdSB-NUjVbR5sKXtgG5MflxeQ-eKNaPJ3lZeMzko7eU2q0
1pCm3vOGlmRfL_keG4t84LItaP3YJdEeMjuuqp0Rq7NWkbwkK-06LuQa2BkSdaE8zdY1944VXumG7esF0gZpzzk9_k50rdDC4HVxQu9DUtizZydzZU_tJlDjivVZ_
3tR1PDWo7vbsWVVOol_yz6cEk-PinO3aaG75GeMHN7HputdzSbziVTTLAu1PXpUnIyWt3gmQBLarMKCCfb7U8vmznnwA5pnXQKwzB9QfkfUkIgA6-007BUBuiRxr-
WGVLY3vNQ87gEEVkdob6k6MJfM9H9Tx2oZ_WtLxiLZ5F2Eh_2P7zNmaqAEkgHQYrw04PzHawL24lTryexA1FdyU
```

Fig. 2 HTTP request header sent during login authentication

during the login sequence are sent to the server via the HTTP request header, unencrypted, as shown in Fig. 2. The censored part of the parameters is the unencrypted password of one of the authors. Any means of hijacking such a request header can prove to be a disastrous flaw within this system.

5.3 *PayTM QC (Quality Control)*

PayTM QC (https://qc.paytm.com) is a quality control dummy website from PayTM which hosts a blatant XSS vulnerability. This vulnerability was fixed at some point in the development cycle of the live website. The vulnerability can be recreated by visiting the website and injecting the following string in the input field of the homepage: *teststring" > < img src = x onerror = alert(document.cookie) >* (Fig. 3).

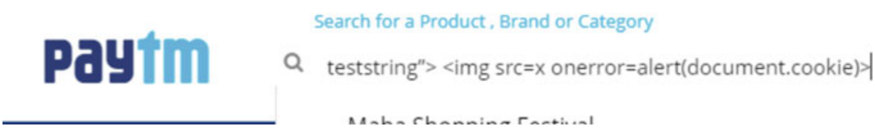


Fig. 3 Input string on the PayTM homepage

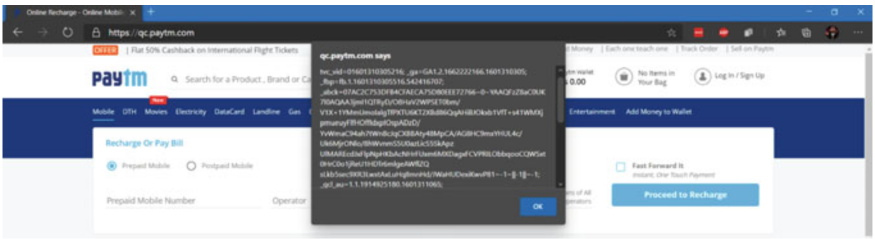


Fig. 4 Alert box upon injection of the input string

The alert box shown in Fig. 4 is a representation of a hijacked cookie. Any attacker could forge a malicious URL using this vulnerability and get access to the victim's personal data using the session cookie.

5.4 *Self-XSS Testing and Results*

Before heading straight to Self-XSS, we discuss a vulnerability we found while listing and conducting preliminary tests on stray targets. One such target was the Navi Mumbai Municipal Corporation (NMMC) Government website (nmmc.gov.in). On testing with a few tags, we found out that the < script > tags which are the most primitive tags used for XSS—were filtered out. However, other document tags such as < h2 > , < a > and < img > were skipped from filtering. Also to note is that the input filter is able to detect event handlers and removes them on rendering the result page. Keeping the above things in mind, we design a relatively simple vector that can act in compliance with our intention: < *a href* = *"this-is-a-test-domain"* > *Click Me* < */a* > (Fig. 5).

Thus, we are successfully able to demonstrate a case of Self-XSS (XSS in which the user copy–pastes the attack vector onto the website himself). This flaw can be exploited in conjunction with social engineering to produce malicious results (Figs. 6 and 7).

Fig. 5 Input vector in the search bar

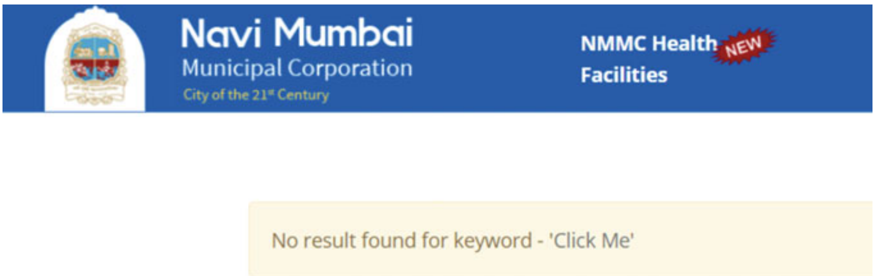


Fig. 6 Result page with the rendered attack vector

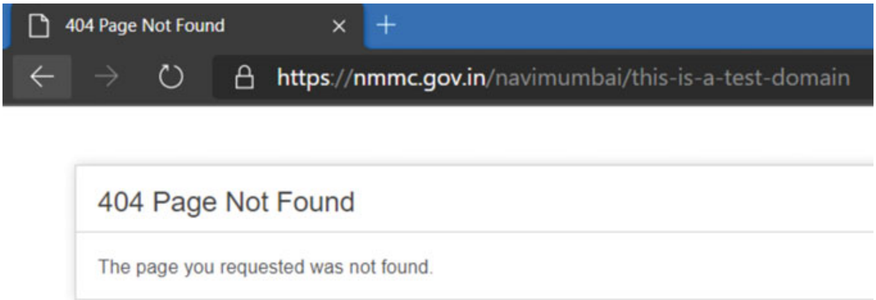


Fig. 7 On clicking the vector element

5.5 *Reflected-XSS Testing and Results*

Reflected-XSS is the type in which the URL contains the attack vector, following which the malicious code is executed on the user end. Anyone that clicks such a link instantly becomes vulnerable to privacy invasion and other threats. This section focuses on the testing of attack vectors onto live websites. An attack vector is a malicious code that is injected into the website to perform XSS. It can range from a simple script tag input to an unsanitized input field—to lengths of base64 strings to bypass WAFs. Web application firewalls (WAFs) prevent certain strings of text from loading the regular intended application and instead load a different application—so that these strings cannot affect the functioning of the website. A list of attack vectors can be found on PortSwigger's website. We tested the following vectors on BigBasket and JioMart: $< xss\ id = x\ tabindex = 1\ onactivate = alert(1) >\ < /xss >$, $< body\ onafterprint = alert(1) >$, $< body\ onbeforeprint = alert(1) >$, $< img\ id = x\ tabindex = 1\ onactivate = alert(1) >\ < /img >$ and $< img\ onafterscriptexecute = alert(1) >\ < script > 1 < /script >$.

At this point in our testing, we realized that the WAF was able to omit critical parts of our vectors—the event handlers (onactivate, onafterprint, onafterscriptexecute and so on) and the alert() string which gives us a confirmation of a successful attack.

Another important thing to note is that certain HTML tags are filtered out by the input applications on these websites, tags such as < script > , < iframe > , which are extremely harmful for an application. However, some tags such as < img > , < marquee > , < svg > —which are common safe components of a website, are allowed on these filters. Thus, we attempt to base our approach around < img > , < marquee > and < svg > tags. We were able to identify that the WAF is triggered when the input detects a regular expression similar to: "onerror = ", "onload = ", "onactivate = " and so on. Basically, when the vector contains an event handler immediately followed by an "equal to" (=) symbol, we are led to the error page. Hence, we attempted to add strings that are accepted as input but bypassed by the input handler. Strings such as !#$%&()* ~ + -_.,:;?@[/|\]^‘. So our vectors are now: < img src!#$%&()* ~ + -_.,:;?@[/|\]^‘ = x onerror!#$%&()* ~ + -_.,:;?@[/|\]^‘ = alert(1) > (Figs. 8 and 9).

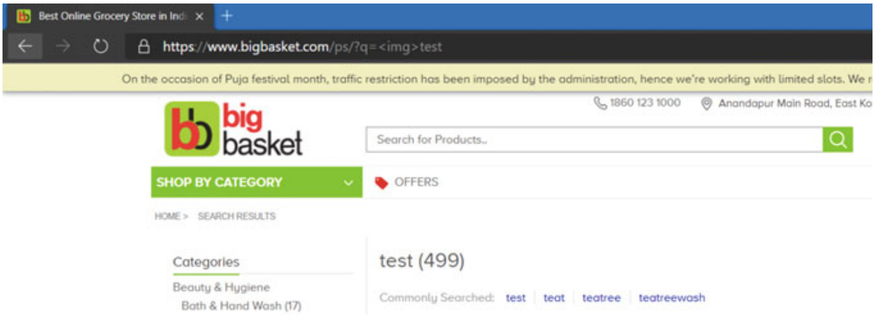


Fig. 8 Result page (notice the lack of < img >)

Fig. 9 WAF gets activated due to the presence of alert() in the attack vector

Moving on from the event handling filter evasion, we move onto the case of bypassing WAF regarding the alert(1) event. This can be done by replacing the command with different methods of representation, or by encoding it in a different way. Other alternative representations to test out are: *(alert)(1), a = alert,a(1),* [1]*.find(alert), top[“al” + “ert”](1), top[/al/.source + /ert/.source](1), a\u0065rt(1), top[‘a\145rt’](1), top[‘a\x65rt’](1)* and *top[8680439..toString(30)](1)* (Fig. 10).

While WAF was not activated upon injecting these vectors, no positive result for the presence of a vulnerability was found. Another way is to encode alert(1) to hex. This approach led to the successful bypassing of filters as shown in Fig. 11.

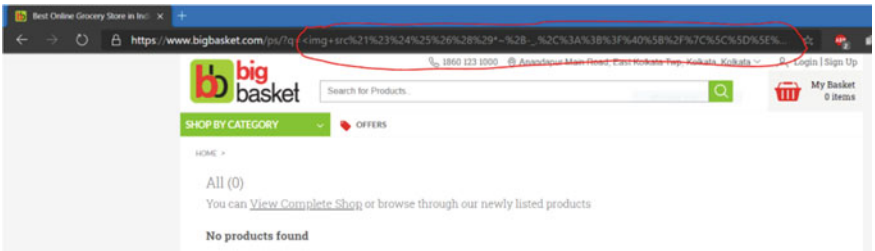


Fig. 10 No WAF activation

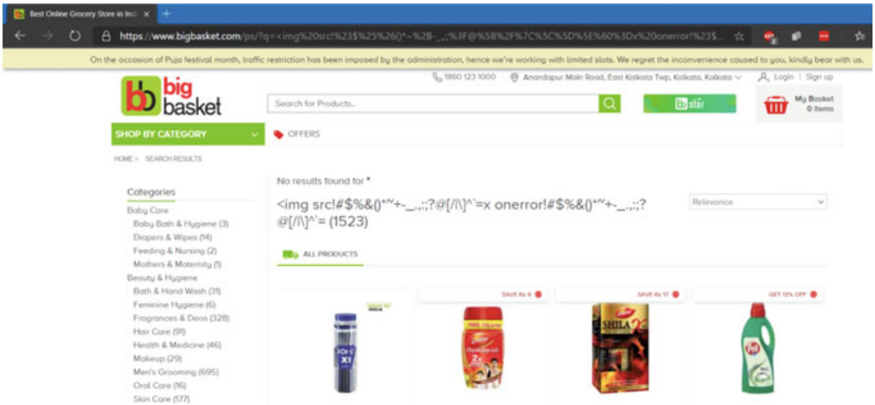


Fig. 11 Script reflected, but no positive result

The results on other targets were either similar or worse. BigBasket seemed to be the most responsive to our vectors, and hence all the testing was focused on it.

6 Conclusion

On researching and testing the possibilities of cross-site scripting on popular web applications, it was found that we were unable to reflect a vulnerability present on the test applications. However, we are able to provide insights to any researcher who intends to continue this effort—while we continue ourselves too. The insights are: 1. BigBasket allows for certain tags to be passed through the WAF: < img > < svg > < marquee > etc. 2. The WAF is able to detect event handlers in their primitive state: "onerror = ", "onload = ", "onactivate = ", etc. 3. BigBasket is an angular-based application. On having further knowledge of angular and the possible XSS vulnerabilities attached with it, one can pursue finding severe attacks on it.

We would finally like to note that XSS is still the most significant vulnerability that prevails in this age, and a lack of research in this field would lead to huge drawbacks leading to security risks and privacy evasion in popular web applications that are in daily use.

References

1. Full-time bug hunting: pros and cons of an emerging career, https://www.helpnetsecurity.com/2020/04/07/bug-hunting-career/
2. Engin K, Christopher K, Giovanni V, Nenad J (2006) Noxes: a client-side solution for mitigating cross-site scripting attacks. In: 2006 ACM symposium on applied computing (SAC '06). Association for computing machinery, New York, pp 330–337

3. Vogt P, Florian N, Nenad J, Engin K, Christopher K, Giovanni V (2007) Cross site scripting prevention with dynamic data tainting and static analysis, NDSS
4. Bisht P, Venkatakrishnan VN (2008) XSS-GUARD: precise dynamic prevention of cross-site scripting attacks. In: Zamboni D (ed) Detection of intrusions and malware, and vulnerability assessment. DIMVA 2008. Lecture notes in computer science, vol 5137. Springer, Berlin, Heidelberg
5. Michael M, Monica SL (2008) Automatic generation of XSS and SQL injection attacks with goal-directed model checking. In: 17th conference on security symposium (SS'08). Association for computing machinery, New York, NY, USA, pp 31–43
6. Sun F, Xu L, Su Z (2009) Client-side detection of XSS worms by monitoring payload propagation. In: Backes M, Ning P (ed) Computer security—ESORICS 2009. ESORICS 2009. Lecture notes in computer science, vol 5789. Springer, Berlin, Heidelberg
7. Weinberger J, Saxena P, Akhawe D, Finifter M, Shin R, Song D (2011) A systematic analysis of XSS sanitization in web application frameworks. In: Atluri V, Diaz C (eds) Computer security—ESORICS 2011. ESORICS 2011. Lecture notes in computer science, vol 6879. Springer, Berlin, Heidelberg
8. Mukesh KG, Govil MC, Singh G (2015) Predicting cross-site scripting (XSS) security vulnerabilities in web applications. In: 2015 12th international joint conference on computer science and software engineering (JCSSE), pp 162–167. IEEE
9. Shailendra R, Pradip S, Jong HP (2017) XSSClassifier: an efficient XSS attack detection approach based on machine learning classifier on SNSs. J Inf Process Syst 13(4):1014–1028
10. A Survey on Cross-Site Scripting Attacks, https://arxiv.org/abs/0905.4850
11. Gupta S, Gupta BB (2017) Cross-Site Scripting (XSS) attacks and defense mechanisms: classification and state-of-the-art. Int J Syst Assur Eng Manag 8:512–530
12. Swaswati G, Nazrul H, Dhruba KB, Jugal K (2017) An unsupervised method for detection of XSS attack. Int J Netw Secur 19(5):761–775
13. Gurpreet K, Bhavika P, Aayushi B, Gargi B, Shashank G (2018) Efficient yet robust elimination of XSS attack vectors from HTML5 web applications hosted on OSN-based cloud platforms. Procedia Comput Sci 125:669–675
14. Gupta S, Gupta BB (2016) XSS-SAFE: a server-side approach to detect and mitigate cross-site scripting (XSS) attacks in JavaScript code. Arab J Sci Eng 41:897–920
15. Qijin W, Jun H, Xiaoxia Q (2019) XSS attack detection and prevention system based on instruction set randomization. IOP Conf Series: Mater Sci Eng 563(4):1–6
16. Germán ER, Jenny GT, Pamela F, Diego EB (2020) Cross-site scripting (XSS) attacks and mitigation: a survey. Comput Netw 166:106960
17. Jingchi Z, Yu-Tsern J, Xiangyang L (2019) Cross-site scripting (XSS) detection integrating evidences in multiple stages. In: 2019 proceedings of the 52nd Hawaii international conference on system sciences, pp 7166–7175, ScholarSpace
18. Shashank G, Brij BG (2016) XSS-immune: a Google chrome extension-based XSS defensive framework for contemporary platforms of web applications. Secur Commun Netw 19(7):3966–3986
19. Burp Suite is the choice of security professionals worldwide, https://portswigger.net/burp

Detecting Cyberbullying with Text Classification Using 1DCNN and Glove Embeddings

R. Sangeethapriya and J. Akilandeswari

Abstract In recent years, the number of social networking sites has exploded, providing a platform for individuals all over the world to connect and discuss their common interests. With the increased use of the Internet and social media platforms, it is no surprise that young people are utilizing these technologies to injure one another, and this happens on a variety of social media sites and apps. In a few of hours, a 10 s Snapchat post can cross platforms and go viral on Facebook, Instagram, and Twitter. The bulk of prior studies have used traditional machine learning models to tackle this challenge, and the bulk of the produced models in these researches are only adaptable to one social network at a time. Deep learning-based models have made their way into the identification of cyberbullying occurrences in recent studies, claiming to be able to overcome the limits of traditional models and increase detection performance. Given the negative effects of cyberbullying on victims, it is critical to identify effective ways to detect and prevent it. Deep learning can aid in the detection of bullies. The main goal is to identify the best algorithm for detecting cyberbullying using 1DCNN, Bidirectional Encoder Representations from Transformers (BERT), and Recurrent Neural Networks (RNNs) (RNN). In comparison with the other two models, 1DCNN and glove embeddings produce more exact results.

Keywords Deep learning · Cyberbullying · Twitter

1 Introduction

Cyberbullying is a widespread issue that affects individuals of all ages and backgrounds. Over half of all teenage people on social media have suffered long term or organized online abuse. As a result, early identification of cyberbullying in

R. Sangeethapriya (✉) · J. Akilandeswari
Department of IT, Sona College of Technology (Autonomous), Salem, Tamil Nadu, India
e-mail: sangeethapriya.r@sonatech.ac.in

J. Akilandeswari
e-mail: akilandeswari@sonatech.ac.in

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_14

social media is critical in reducing the effects on victims. According to the most recent research on bullying, more than one-third of students aged 13–15 are bullied, and nearly three-tenths of teenagers in 39 European and North American countries confess to bullying others at school. Cyberbullying is the act of intimidating, insulting, or targeting another person through the use of electronic communications, such as pictures or videos. Cyberbullying takes place on a variety of social media sites, including chat rooms, blogs, and instant messages. Cyberbullying has the potential to cause significant damage because it leaves a lasting mark in cyberspace and can easily reach a large audience. Bullying and cyberbullying reinforce each other, resulting in a vicious cycle of negative conduct 13. Bullying is a major concern for children, whether it occurs online or in person. Despite the fact that rates vary by region, it is present and prevalent around the world, affecting a large number of children as victims, offenders, or bystanders.

When asked which social media platform they use the most, 52.80% (n = 87) said they use Facebook, 27.95% (n = 46) said Instagram, and 19.25% (n = 32) said Twitter. Sixty-six percent of online abuse victims said they were threatened on Facebook. Around 37% of children aged 12 to 17 have been the victim of cyberbullying at least once. According to Statista, 38% of online trolls tend to harass people on social media, while 23% prefer to harass people on YouTube and other video-sharing sites. They are also involved on message boards, chat rooms, and blogs.

Initiatives to combat cyberbullying and cybersafety are only now being developed and tested. Websites, advice sheets, and other Internet resources may provide parents with information on how to effectively secure their children [1]. Furthermore, it shows that these Internet services are frequently advertised by product-selling groups and are rarely scientifically based. There is a risk that this information will be detrimental if it is not supported by research [2]. Parents, educators, school leaders, and medical professionals must exercise caution when evaluating material on these sites and instead depend on online resources offered by government entities and activist groups that utilize statistics to drive their suggestions.

Cyberbullying, traditional bullying, and mixed bullying all had incidence rates of 9.9, 13.3, and 9.4%, respectively, implying that one among students were victims of one of these forms of bullying; 48.7% of cyberbullying victims also endured traditional bullying, and 41.5% of traditional bullying victims also experienced cyberbullying.

1.1 Types of Cyberbullying

Verbal Bullying

Verbal bullies use gestures, comments, and name-calling to acquire control and power over their victims. Verbal bullies, are bullies who use their words to ridicule, degrade, and hurt others [3]. They select their victims based on their appearance, actions, or behavior. Children with specific disabilities are also harmed by verbal bullies.

Because assaults almost always occur while adults are not present, verbal abuse can be difficult to detect. As a result, one individual's word is frequently set over another's. Furthermore, many conservatives feel that what youngsters say has little bearing on how others perceive them. As a result, they generally tell the bullied individual to "ignore it." On the other hand, verbal abuse should be treated seriously.

Relational Aggression

Interpersonal violence is a complex and subtle kind of bullying that many parents and teachers fail to recognize. Relational violence, also known as emotional abusive behavior, is a type of social coercion in which tweens and teenagers seek to hurt or undermine the social standing of their friends.

Bullies in relationships frequently remove people from groups, propagate stories, take advantage of circumstances, and undermine trust [4]. The goal of a prosocial behavior bullying is to raise their social standing through manipulating or harassing another individual. In general, girls utilize relational aggressiveness more than boys, especially in between fifth and eighth grades. These girls are referred to as "evil girls" or "frenemies" at times. An adolescent or preteen who is the victim of relational violence is likely to be mocked, humiliated, dismissed, excluded, and threatened.

Tweens are not the only ones who experience emotional violence in middle school. Relational violence is used by some bullying bosses and other office bullies.

Bullying in the traditional sense is normally restricted to specific times and locations, such as the playground or the school bus. This gives the target a sense of predictability, and he or she can feel safe at certain times and in certain locations [5]. Technology, on the other hand, is everywhere, and we are constantly surrounded by our phones and computers. Cyberbullying can occur at any time of day or night, seven days a week, or 356 days a year. It can happen right next to us or thousands of miles away. Finding a secure location is difficult for the goal.

When a tween or adolescent uses the Internet, a cellphone, and perhaps other technologies to harass, threaten, shame, or threaten some other person, this is referred to as cyberbullying. When an adult is involved in the harassment, it is referred to as cyber-harassment or cyberstalking.

Cyberbullying includes posting unpleasant pictures, making online threats, and sending offensive messages via email. Cyberbullying is on the rise among teenagers and tweens, owing to their constant access to the Internet. Bullies can bully their targets with even less chance of being detected, so it is becoming more common.

Cyberbullies will occasionally say things they would never say in person. Because of technology, they felt anonymous, isolated, and removed from the issue. Victims of cyberbullying describe it as invasive and have never. Bullies will target individuals at any time and from any location, along with their own residences. As an outcome, bullies can have far-reaching consequences.

2 Related Work

Perera et al. [6] proposed a technique for identifying cyberbullying on Twitter that combines TF-IDF, N-gram, profanity, and text analytics. TF-IDF and common words were used to determine the relevance of terms in a document, and the author looked for a curse word there with a subject in the text. The authors of this paper emphasize the importance of standardizing data in order to achieve high efficiency, so they created a module called Content-based Features and Sentiment-based Features to delete non-essential words from tweets while keeping emoticons and necessary details. The proposed approach outperforms all current methods in terms of 74.50% accuracy, 74% precision, 74% recall, and 74% F1-score as compared to machine learning and other deep learning methods.

Sangwan et al. [7] used a swarm-based wolf search algorithm to detect cyberbullying messages on the PHEME dataset. The inputs are pre-processed, and TF-IDF is used to extract functionality. The wolf search algorithm is used to formulate the results of feature extraction. The model measures the visual range's current location and detects the cyberbullying post. When the method's results are compared to an established decision tree classification model, it is clear that WSA has a higher accuracy percentage for both training and test datasets.

Nikhila et al. [8] suggested the use of a GAN (Generative Adversarial Neural Network) with Glove to detect cyberbullying messages. YouTube, FormSpring, and Myspace provided the input datasets. Glove is used to perform word embeddings. This study employs two neural network architectures. The first is the Generator network, which takes a batch of sentences as input and outputs an equal-dimensional vector for each sentence [9]. The real oversampled data is the performance of this network. The Discriminator network is another architecture that is used to separate the original data from the output of the generator network and perform binary classification. The output of this network, along with the corresponding class labels, is appended to the original data. Both networks will be trained together in this way, and the produced data will be reliable for oversampling. Glove outperforms other word embeddings such as CBOW and Word2vec, according to the results. As a result, this model is more effective, and the results are more precise.

Cyberbullying identification in social media was introduced by Yadav et al. [10] and Aija [11] because cyberbullying on online channels is a difficult task because it occurs in a variety of ways, including the use of images, toxic messages, and videos, among others. To solve the problem, the author created a model that uses BERT to detect cyberbullying messages. The input dataset is obtained from spring, and it is pre-processed before being fed into the BERT-pre-trained model, where it is converted to integer tokens and compared to previously specified values, after which the classifier layer is used to predict the bully message. As a machine learning model, the baseline model is used. When the results of the BERT and machine learning models are compared, it is clear that the proposed model has a higher accuracy percentage of 98%.

LSHWE (Locality-Sensitive Hashing word embedding) is a technique used by Zhao et al. [12] to detect cyberbullying. Glove and word2vec have been used in the past to insert terms. However, such architecture is based on uncommon terms with little meaning. The authors propose using LSHWE in conjunction with Nearest Neighbor Search to resolve this issue (NN). The authors say that this method represents uncommon words well and cuts down on running time. To learn word representations, this architecture employs an autoencoder model. When the LSHWE, Glove, Word2vec, and Sentiment-specific word embeddings (SSWE) results are compared, the LSHWE is found to be more accurate.

To detect bully posts, Wang et al. [13] used Multimodal cyberbullying detection (MMCD). Instagram and vine provided the datasets. This is accomplished by the use of BiLSTM and a hierarchical focus mechanism. The authors believe that by using this model, CNN would be able to solve two major flaws: (1) It does not work well when the text lacks sequence detail and (2) concentrates on a single modality. The authors tested two datasets, Vine and Instagram, with two pre-trained word vectors, Word2Vec and Glove. Glove is found to be more powerful than the other two word vectors, demonstrating that the MMCD and BiLSTM algorithms outperform the baseline model of machine learning algorithms by 83.8 and 84.4%, respectively.

Vimala Balakrishnan and colleagues (2020) aimed to use a dataset of 5453 tweets with the hashtag, tweets, and retweets, and user attributes such as identities, sentiment, and emotion were used to enhance cyberbullying identification. Features are derived from the tweets after they have been pre-processed. The authors tested the baseline SVM, Random Forest, J48, and Naive Bayes models. For the cyberbullying classifications, both Random Forest and J48 performed admirably.

Fortunatus et al. [14] suggested a lexicon-improved rule based on Facebook comments to detect cyberbullying messages. The data is pre-processed before being translated to a normalized vector. Swearword lexicon, text violence, text positive word, and text sentiment score all process the normalized vector's performance. The sentence is graded as a bully post or not based on this ranking. The best result was obtained with an F1-score of 86.673% and a Recall of 95.981%, according to the writers.

Lopez-Vizcaino et al. [15] aimed to detect cyberbullying messages using an early detection approach based on text similarities and time features. Vine provided the datasets. To begin, the Random Forest classification algorithm is used to categorize the comments. Second, based on their threshold function, the threshold values are determined and the posts are classified. On the other hand, the Extra Tree approach (ET) is used to identify text messages, with a threshold value varying from 0.9 to 0.5. The system outperformed the Support Vector Classifier and Logistic Regression baseline models (LR).

For detecting violence on tweets from the cyber-troll dataset, Sadiq et al. [16] used TF-IDF, unigram, and bi-gram. The proposed system evaluated the efficiency of MLP, CNN-LSTM, and CNN-Bi-LSTM by manually extracting and selecting important features and completely linked layers of neural networks. Their best recorded results were 92% accuracy, 90% precision, 90% recall, and 90% F1-score, respectively (Table 1).

Table 1 Analysis of studies on cyberbullying detection in a comparative manner

Author	Year	Dataset	Data feature	Model used	Metric used	Short comings
Andrea Perera et al.	2020	Twitter	TF-IDF values, N-gram, and profanity	SVM classifier	74.50% accuracy,74% precision, 74% recall, and 74% F1-score	Limited dataset and features only used
Saurabh Raj Sangwan et al.	2020	PHEME	TF-IDF and bag-of-Word	Swarm-based wolf search algorithm (WSA)	Accuracy 95%	Domain dependent, text information only used
Munipalle Sai Nikhila et al.	2020	FormSpring and Twitter	Glove word embeddings	Generative adversarial Networks	F1-score of 0.81	Binary classification method only used to identify cyberbullying message
Jaideep Yadav et al.	2020	Form spring and Wikipedia	word2vec	BERT	Accuracy—96%	Computational complexity is high
Zehua Zhao et al.	2020	Twitter	Locality-sensitive hashing word embedding	Nearest Neighbor Search	F1-score 0.8629	Only small dataset only used
Kaige Wang et al.	2020	Vine and Instagram	Word2vec	multimodal cyberbullying detection (MMCD)	F1-score 0.841	Text only used for classification
Vimala Balakrishnan et al.	2020	Twitter	Twitter API	Random forest and J48	Accuracy 91.88%	Limited features only used such as popularity and number of followers

(continued)

Table 1 (continued)

Author	Year	Dataset	Data feature	Model used	Metric used	Short comings
Meisy Fortunatus et al.	2020	Facebook	TF-IDF	Lexicon enhanced ruled based method	Recall 95.981%	Computational complexity is high
ManuelF.Lopez-Vizcaino et al.	2021	Vine	BOW	extra tree model	Accuracy 92%	Effective techniques are required to calculate accuracy. Classifier performance depends on only number of posts
Saima sadiq et al.	2020	Twitter	TF-IDF	CNN-LSTM	Accuracy 92%	Limited features only used

Methodology

The data description and models used in the experiments are discussed in this section. The first subsection goes over the specifics of the data we used in our experiments, while the second goes over the proposed method.

Dataset Description

To train and test the proposed model, we used the Kaggle dataset, which was collected from various social networks (Twitter, Facebook, etc.). A total of 35,000 posts are included in this dataset. Comment and comment mark were included in the dataset.

3 Proposed Approach

We go through a few baseline approaches before moving on to the proposed solution. Linear SVM, Logistic Regression, TF-IDF, Grid scan, and LSTM are examples of baseline methods.

The following are the comparison experiments:

Logistic Regression + Char n-gram: The experiment uses character-based embedding for the term vector.

Random Embedding with LSTM: In the experiment, the LSTM and its internal memory were utilized to analyze random sequences of inputs in order to catch long-term relationships in tweets that might be beneficial in hate speech identification.

The SVM unigram is as follows: The traditional method employs 70% training and 30% testing results, as well as the NLTK vectorization method. In this experiment, a single unigram word sequence is used.

CNN: Word2vec is used as the word embeddings in the standard convolutional neural network structure. Padding is used in the convolution operation, and all the sentences are the same length.

LSTM + Random Embedding + GBDT: Texts are transformed to random values, then an LSTM is generated utilizing backpropagation, as well as the obtained embeddings are utilized to train a GBDT class.

Word Embedding

Word embedding captures more detailed features and textual content involved in human language, as well as semantic and syntactic connections between concepts. Word embeddings are a type of representation that allows concepts with comparable meanings to be represented. People often use plain to show their aggression, for example. Word2vec, an unsupervised word embedding-based method, CBOW, is used to find semantic and syntactic word relationships. Word2vec makes use of a fully connected neural network with only one hidden layer.

FastText, an extension of the word2vec model, is another technique to word embedding. Rather of learning vectors for words directly, FastText represents every text as an n-gram of letters. This enables the embeddings to understand suffixes and prefixes as well as the meaning of shorter words. A term frequency-inverse document frequency (TF-IDF) is a term frequency-inverse document frequency. It aids in determining the relative meaning of a given word in relation to other words in the text and corpus. It measures TF and IDF in two different ways. When you combine the two, you will get a TF-IDF score.

To effectively achieve the context of the data available, we construct a function with Glove. It is a variant of the Pennington et al. Stanford created word2vec technique for fast learning word vectors. Glove is a method for integrating statistical data derived from matrix factorization methods such as LSA and local context-based learning derived from word2vec.

Feature Vector Construction

The inputs are first pre-processed, with stop terms, URLs, and punctuations omitted. The pre-processed text is converted into numbers using glove word embedding techniques, with each word represented by an array of numbers. The co-occurrence probability ratio between two terms is represented as vector differences. Glove's goal is to decrease the ratio between the dot function of two words' vectors and the logarithm of the number of times they appear together.

$$\text{G} = \sum\nolimits_{q,r=1}^{x} b[y_{q,r}](c_q^v \tilde{c}_r + e_q + \tilde{e}_r - \log\ \text{b}_{q,r}) \tag{1}$$

where c_q and e_q represent the vector representation and bias of word q, $\widetilde{c_r}$ and e_q represent the background word vector and bias of word n, $b_{q,r}$ refers to the number of periods word n appears in the context of word r, and b is a weighting function that gives lower weights to uncommon and regular co-occurrences.

The training aim of Glove seems to be to learn vector representation with a dot product equal to the exponential of the co-occurrence likelihood of the words. Because the logarithm of a ratio equals the difference of logarithms, this objective connects ratios of co-occurrence rates with vector differences in the word vector space. Because these ratios can convey some meaning or significance, this information is likewise stored as vector differences. As a result, the resultant word vectors excel in word analogy tasks.

3.1 Cyberbullying Detection Using 1DCNN

We then proceed to detect instances of cyberbullying in Twitter media sessions using the labeled ground truth info. The data was subjected to five-fold cross-validation in order to build and train the classifier. The dataset is divided into k equal-sized sets, with one set used for testing and the other partitions for training. This allows k separate runs to be performed, with each partition serving as a testing set once. As a result, the higher the k, the more precise the model evaluation. In each run, 80% of the data was used for training, while 20% was used for testing. A word embedding can use a convolutional kernel. These kernels will slide down a list of word embeddings in order to process an entire series of terms. A single kernel will look at the first word embedding, then the next word embedding, and so on, moving down a list of input embeddings one by one. The final output will be a function vector of about the same number of values as the number of input embeddings, so the size of the input sequence does matter. The ReLU classifier classifies texts using the dense layer, max pooling layer, dropout layer, and output layer.

The formula for a one-dimensional convolutional layer is:

$$y_q^p = \mathrm{f}(\sum\nolimits_p^N y_p^{l-1} * \mathrm{M}_{pq}^l + \mathrm{a}_q^l) \tag{2}$$

where M is the number of convolution kernels, q is the number of kernels, N is the channel number of the input y_p^{l-1}, a is the kernel's bias, f () is the activation function, and * is the convolution operator. Convolution should be performed on the dataset with filters of 250, kernels of 3, and secret dimensions of 250. We also use a chronic dropout with p = 0.2 to prevent overfitting. Since the outputs of a layer under dropout are randomly subsampled, the capacity of the network is reduced or thinned during training. Finally, to avoid the vanishing gradient problem and boost accuracy, a completely connected layer, namely a dense layer with a ReLU activation function, was used.

3.2 Recurrent Neural Network

RNNs are feedforward networks that have been generalized to model sequential data. It takes an input and produces an output right away. It is carried out using a loop mechanism. Glove word embeddings transform the inputs to word vectors. RNN processes the output of word embeddings. It moves the previous secret state to the next stage of the sequence while processing. The secret state serves as the internal memory of the neural network. It contains details about previous data that the network has seen. To begin, a vector is created by combining the input and previous hidden state. The current and previous inputs are now represented in that vector. The vector is enabled by ReLU, and the result is the new hidden state.

RNNs keep track of their state over time, allowing them to process feedback of varying lengths. When it comes to hate speech of various lengths, to determine whether a post is cyberbullying or not, a word will be used as a token, and the resulting score from previous states will be used as input to the current state. Finally, ReLU is employed to prevent the issue of disappearing gradients.

The current state formula is as follows:

$$k_t = g(k_{t-1},\ y_t) \tag{3}$$

The new state is k_t, the previous state is k_{t-1}, and the existing state is y_t.

The network receives the input and then calculates its current state by combining the current input with the previous state. For the next time step, the current k_t becomes k_{t-1} and can go as many time steps as the problem requires, combining the knowledge from all previous states. The final current state is used to measure the performance y t after all of the time measures have been completed. After that, the output is compared to the real output, and an error is calculated. The error is then backpropagated to the network, which updates the weights and trains the network.

3.3 Bert

Bidirectional Encoder Representations from Transformers: Transformer is made up of two different mechanisms. The text input is interpreted by an encoder, and the text is decoded to make a prediction. The transformer encoder reads the entire word series in one go. As a result, it is classed as bidirectional; however, it would be more accurate to call it non-directional. This feature allows the model to infer that meaning of a word based on its surrounds.

Before being processed by the neural network, the input consists of a series of tokens that are wrapped into vectors. After applying a classification layer to the encoder output, the output vectors are multiplied by the additive to transform them to the lexical dimension. Softmax is used to determine the probability of every word in the vocabulary to avoid overfitting.

Result and Discussion

We equate our results to those of the LSTM baseline model, Charn gram logistic regression, SVM, and CNN Comparison Experiment. Cyberbullying identification is assessed using performance metrics such as accuracy, precision, recall, and F1 ratings. The uncertainty matrix is the foundation for all experimental findings. The criteria that we concentrated on are mentioned below [17]. Different architecture 1DCNN and Glove, BERT, RNN, and machine learning algorithms are used in the experiments. 1DCNN architecture was used, with Glove embedding (200 embedding layers, 32 batch size), 1D max polling layer, and ReLU layer. To avoid the issue of overfitting, a dropout layer was used, with a filter size of 5 and secret units of 250. For the hate speech data collection, the following tests were carried out, and the applied strategies were compared.

Confusion Matrix

Predictive value	Actual value	
	Positive	Negative

(continued)

(continued)

	True positive (TP)	False negative (FN)
	False negative (FN)	True negative (TN)

2. Accuracy score: $\frac{\mathrm{TP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$
3. Precision: $\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$
4. Recall: $\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$
5. F1-score: $\frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$

We may discover a few accuracy-based criteria for further investigation, such as precision, recall, and F conclusion. Precision is defined as the proportion of similar instances found between the retrieved instances. Recall is the division of recovered complementary instances into the total number of relevant instances, whereas extract is the division of retrieved complementary cases into the total number of specific instances. Precision and recall are also properly depending on importance of comprehension and design. The harmonic mean of precision and recall, also known as the standard F-measure or balanced, is a measure that combines precision and recall.

Iterations, batch size, and epoch: The number of times an evaluated dataset is transported forward and reverse through the neural net ONCE is specified as an Epoch. We break such an epoch into multiple smaller batches since it is too hefty to feed to the computer all at once. The number of iterations is the number of batches expected to accomplish one epoch (Table 2).

The table summarizes that CNN produces a better result than SVM. To provide a better comparison view, Fig. 1 shows the graphical comparison of CNN and SVM. Figure 2 compares precision between SVM and CNN. Finally, Fig. 3 graphically compares recall of SVM and CNN (Figs. 4 and 5).

Table 2 Comparison of accuracy with various number of epochs

Method	No. of epochs	Loss	Accuracy
1DCNN	1	0.1470	0.9487
	2	0.1220	0.9567
	3	0.1105	0.9578

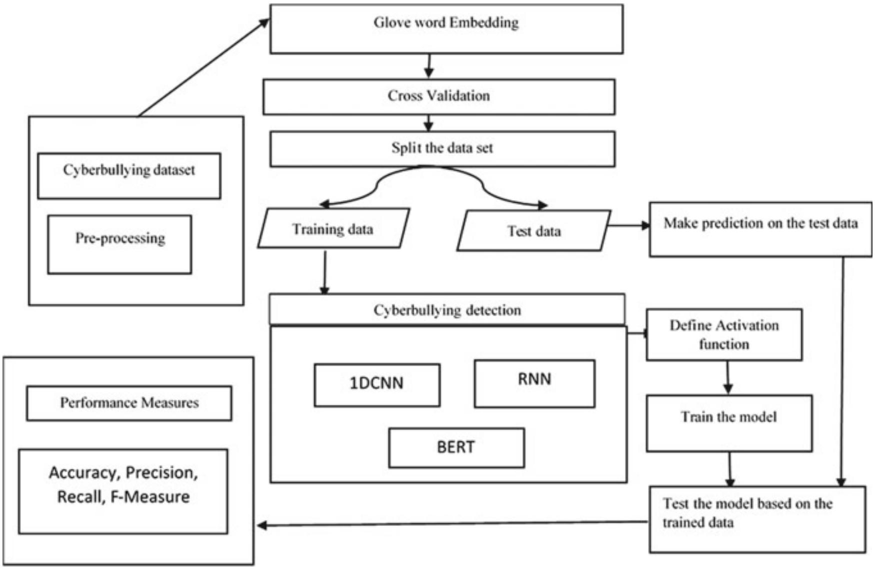


Fig. 1 Proposed system architecture

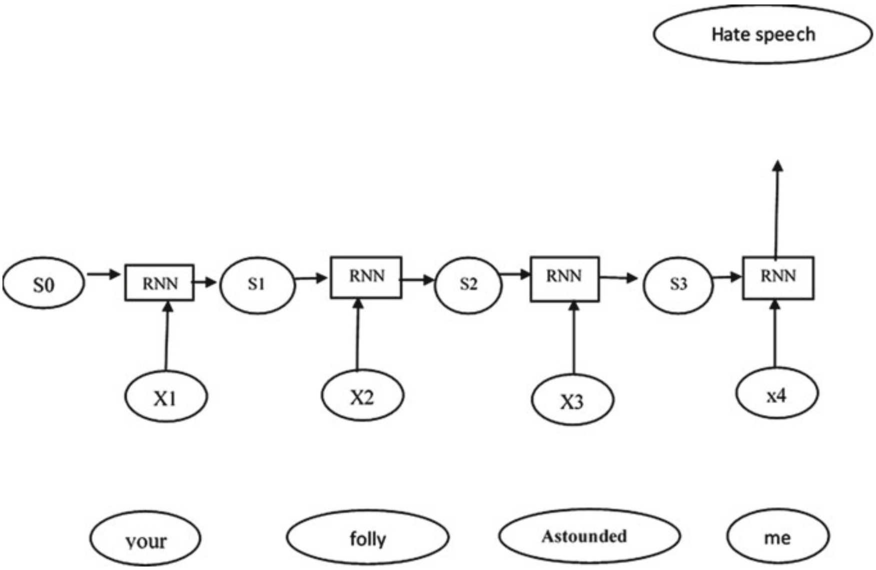


Fig. 2 Text classification using RNN

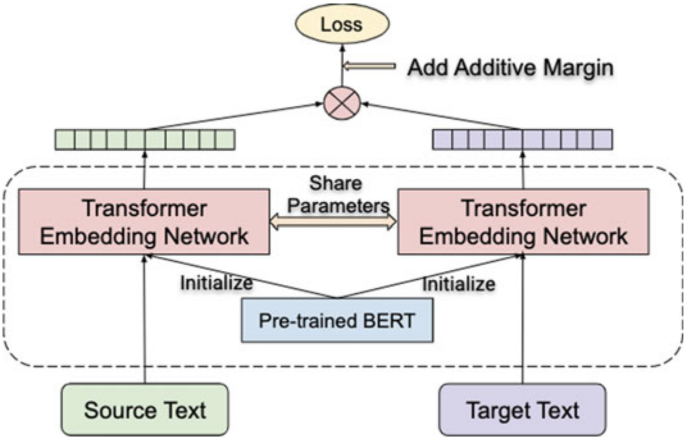


Fig. 3 Architecture of BERT

Fig. 4 Comparison of loss and epochs

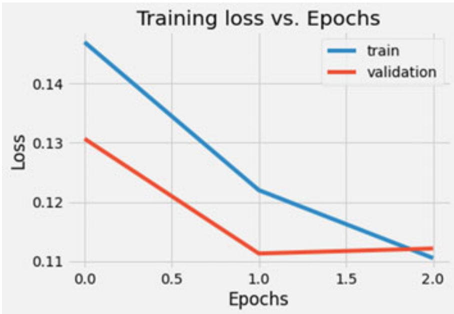
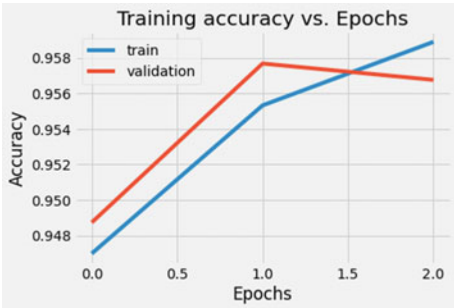


Fig. 5 Comparison of accuracy and epochs



Glove word embedding is based on global word co-occurrence counts and makes good use of statistics by minimizing least-squares error and generating a word vector space as a result. With vector distance, such an outline adequately preserves word similarity (Figs. 6, 7, and 8) (Tables 3 and 4).

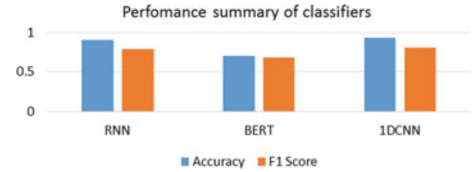


Fig. 6 Performance summary of classifiers

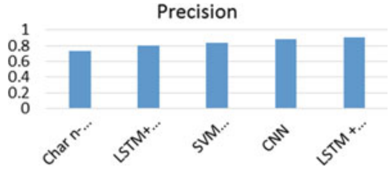


Fig.7 Comparison between baseline methods in terms of precision

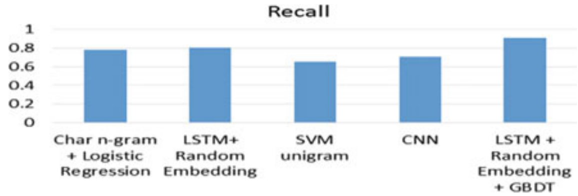


Fig. 8 Comparison between baseline methods in terms of recall

Table 3 Algorithms of classification comparison

Classifier model	Accuracy	F1-score
RNN	0.910	0.79
BERT	0.698	0.687
1DCNN	0.931	0.813735

4 Conclusion and Future Work

In this paper, we present a comparison of machine learning and deep learning classifiers for detecting cyberbullying in Twitter postings. The classifiers were trained using a collection of Twitter text messages. The 1DCNN and glove embeddings for text classification were proposed; the SVM, logistic regression, LSTM, BERT, and

Table 4 Comparison of baseline method of classification algorithms

Model	Precision	Recall	F1-score
Char n-gram + logistic regression	0.729	0.778	0.753
LSTM + random embedding	0.805	0.804	0.804
SVM unigram	0.8353	0.65	0.840
CNN	0.883	0.7028	0.9023
LSTM + random embedding + GBDT	0.903	0.903	0.903

RNN were compared with IDCNN in which 1DCNN performs with glove word embeddings with better accuracy. 1DCNN can efficiently learn the key characteristics from raw data, and just one-dimensional convolutional is done, resulting in a simpler structure and fewer parameters, using less computer resources and time and resulting in low-cost applications owing to their minimal computational needs. The future work can be building a system which will detect cyberbullying content on image data and identifying the bullying user profiles.

References

1. Chan TKH, Cheung CMK, Lee ZWY (2021) Cyberbullying on social networking sites: a literature review and future research directions. https://doi.org/10.1016/j.im.2020.103411
2. Ali WNHW, Mohd M, Fauzi F (2018) Cyberbullying detection: an overview, 978-1-5386-7541-0/18/$31.00 ©2018 IEEE
3. Sanchez-Medina AJ, Galvan-Sanchez I, Fernandez-Monroy M (2020) Applying artificial intelligence to explore sexual cyberbullying behaviour. https://doi.org/10.1016/j.heliyon.2020.e03218
4. Chatzakou D, Leontiadis I, Blackburn J, De Cristofaro E, Stringhini G, Vakali A, Kourtellis N (2019) Detecting cyberbullying and cyberaggression in social media. ACM Trans Web 13(3). https://doi.org/10.1145/3343484
5. Aind AT, Ramnaney A, Sethia D (2020) Q-Bully: a reinforcement learning based cyberbullying detection framework, 978-1-7281-6221-8/20/$31.00 ©2020 IEEE
6. Pereraa A, Fernandob P (2021) Accurate cyberbullying detection and prevention on social media. Procedia Comput Sci 181:605–611. https://doi.org/10.1016/j.procs.2021.01.207
7. Sangwan SR, Bhatia MPS (2020) Denigration bullying resolution using wolf search optimized online reputation rumour detection, pp 305–314. https://doi.org/10.1016/j.procs.2020.06.036
8. Nikhila MS, Bhalla A, Singh P (2020) Text imbalance handling and classification for cross-platform cyber-crime detection using deep learning
9. Espelage DL, Hong JS (2017) Cyberbullying prevention and intervention efforts: current knowledge and future directions 62(6):374–380. https://doi.org/10.1177/0706743716684793
10. Yadav J, Kumar D, Chauhan D (2020) Cyberbullying detection using pre-trained BERT model. ISBN: 978-1-7281-4108-4, 978-1-7281-4108-4/20/$31.00 © IEEE pp 1096–1100
11. Aija, Protecting children from bullying, https://violenceagainstchildren.un.org/news/celebrating-childhoodjourney-end-violence-against-children
12. Zhao Z, Gao M, Luo F, Zhang Y, Xiong Q (2020) LSHWE: improving similarity-based word embedding with locality sensitive hashing for cyberbullying detection, 978-1-7281-6926-2/20/$31.00 © IEEE
13. Wang K, Xiong Q, Wu C, Gao M, Yu Y (2020), Multi-modal cyberbullying detection on social networks, 978-1-7281-6926-2/20/$31.00 ©2020 IEEE
14. Fortunatusa M, Anthonya P, Chartersa S (2020) Combining textual features to detect cyberbullying in social media posts. Procedia Comput Sci 176,pp 612–621. https://doi.org/10.1016/j.procs.2020.08.063.
15. López-Vizcaíno MF, Nóvoa FJ, Carneiro V, Cacheda F (2021) Early detection of cyberbullying on social media networks. Futur Gener Comput Syst 118:219–229. https://doi.org/10.1016/j.future.2021.01.006
16. Sadiq S, Mehmood A, Ullah S, Ahmad M, Choi GS (2021) Aggression detection through deep neural model on Twitter. Future Gener Comput Syst 114:120–129. https://doi.org/10.1016/j.future.2020.07.050

17. Cuzcano XM, Ayma VH (2020) A comparison of classification models to detect cyberbullying in the Peruvian Spanish language on Twitter 11(10):132-138
18. Wang CW, Musumari PM, Techasrivichien T, Suguimoto SP, Tateyama Y, Chan CC, Ono-Kihara M, Kihara M, Nakayama T (2019) Overlap of traditional bullying and cyberbullying and correlates of bullying among Taiwanese adolescents: a cross-sectional stud. https://doi.org/10.1186/s12889-019-8116-z

A Bayesian Network-Based Software Requirement Complexity Prediction Model

Halima Sadia, Syed Qamar Abbas, and Mohammad Faisal

Abstract Software development has an inflated probability of project failure and the major reason for it is the poor requirement engineering process. Potential threats or risks related to requirements must be identified at the earlier stages of the development itself, so as to minimize the negative impact of subsequent affects. Researches reveal that VUCA risks, i.e., Requirement Volatility, Requirement Uncertainty, Requirement Complexity, and Requirement Ambiguity, are the basic sources of risks for other risks too. Complexity in requirements is one of the important factors affecting quality of the product. Computing and analysis of the product complexity in the requirement analysis phase of SDLC will give benefits in assessing the required development and testing efforts for the prospective software product. Failing to which, software designers and testers will need further clarification, thus slowing down the design and verification process. This paper attempts to establish a connection between the VUCA risks and propose a methodology to minimize requirement complexity. The various factors affecting requirement complexity are identified, in the requirement engineering phase. A Bayesian approach is proposed to predict the requirement complexity. The proposed model uses various complexity factors found through extensive literature review to manage requirement complexity of the software products.

Keywords Requirement complexity · Bayesian network · Software development · Software requirement risk

H. Sadia (✉)
Department of Computer Science &, Engineering Integral University, Lucknow, India
e-mail: halima@iul.ac.in

S. Q. Abbas
Ambalika Institute of Technology & Management, Dr. A. P. J. Abdul Kalam Technical University, Lucknow, India

M. Faisal
Department of Computer Application, Integral University, Lucknow, India

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_15

1 Introduction

During the elicitation phase, if the requirements have not been captured successfully, the entire development technique will fail leading in a loss of time and financial costs [1]. Understandable requirements are crucial for software design process. However, in practice it is difficult to obtain the preferred degree of understandability, because in big software products a significant number of requirements have a tendency to have ambiguous or complex decisions.

Developers usually begin with requirements that are ambiguous, uncertain, incomplete, insufficient, and volatile in nature. With the passage of time, such requirements slowdown the development speed and increase the danger of late design adjustments, consequently finding and improving them is an urgent mission for software developers and designers [2–5]. With time, the requirements evolve, resulting in addition, deletion and/or modification of functionalities in order to achieve a quality product [6–9].

In order to empower requirement complexity prediction, this research work has been conducted with the objective to introduce an efficient technique to predict requirement complexity at an early stage.

2 Requirement Complexity

In the past few decades, software complexity has always created new horizons for researchers. Software complexity can be described as the primary driving force of cost, and overall performance of software products.

Nevertheless, there is no collective clearance on the definition of software complexity; however, most of the views are agreed on Zuse's view of software complexity [10] "software complexity is the degree of difficulty in analyzing, maintaining, testing, designing and modifying software".

Software complexity is a problem that persists within the whole software development method and affects each and every stage of the software lifecycle [11].

Requirement complexity can be described as the degree to which a requirement is difficult to understand and verify [12, 13]. There are various factors that contribute to the level of complexity and make each requirement a more or less complex. Requirement Complexity Approaches are briefly discussed in following Table 1.

3 Proposed Requirement Complexity Prediction Model

This paper proposes a model for measuring the requirement complexity based on the attributes which contributes to defining the software requirements. These attributes have been derived through extensive literature review and also on the basis of the

Table 1 Relevant work requirement complexity management

Study	Title	Contribution
Sharma, Kushwaha [14]	A complexity measure based on requirement engineering document [14]	It identifies complexity of software straightaway freezing and fixing the requirements Cognitive information and code complexity measures are used
Sharma, Kushwaha [15]	Applying requirement-based complexity for the estimation of software development and testing effort [6]	This paper presents an approach to predict testing efforts using improved requirement-based complexity (IRBC)
Purawinata [16]	The prediction of software complexity based on complexity requirement using artificial neural network [7]	The work presents a model between prediction complexity of software and requirement complexity using artificial neural network method

recommendations of IEEE: 830:1998 for software requirements to write Software Requirements Specification (SRS) document [14–16]. Identification and classification of the complexity attributes is the preliminary requirement for the development of the proposed model.

It has been revealed that the basic components of the requirement like number of requirements, number and types of stakeholder, etc., make the requirement complex [17]. These basic components have been classified as follows (Fig. 1).

1. Requirements and Expectations
2. Parties Involved
3. Development Characteristics
4. Development Team

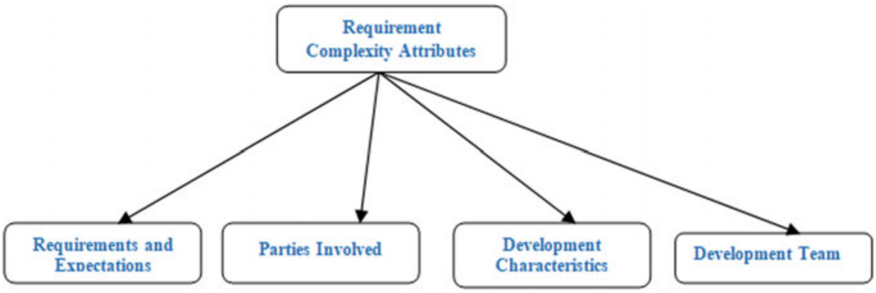


Fig. 1 Requirement complexity attributes

3.1 Requirements and Expectations

Requirement documentation is a vital step in requirement engineering process. The quality and complexity of requirements is based on the number of requirements and interdependencies among them. Every requirement can be partitioned into following subset of attributes that comprise the complexity of a requirement [18–22].

- Total No. of requirements
- Interdependency among requirement
- Conflicting Requirement Objectives
- Clear Statement of Requirements (Requirement Ambiguity)
- Uncertain Requirements
- Changing and Regulatory Requirements

3.2 Parties Involved

The requirement of a software product is being defined by various stakeholders. These stakeholders may vary in different types like their culture and technical knowledge about the domain. These factors affect complexity of a requirement and thus require attention while measuring the requirement complexity. Literature reveals that user involvement is one of the main factors that affect a successful project development. Different sub-attributes may include the following:

- No. of Stakeholder
- Categories of Stakeholders
- Cultural Variety
- Geographic Distance
- Stakeholders Technology illiteracy
- User Involvement

3.3 Development Characteristics

The development environment has a great impact on the software product being developed. Often, the environment challenges to the system under development and as a consequence the requirements have to be adjusted accordingly and sometimes resulting in increased complexity of requirements. Various sub-factors are as follows:

- Number of Interfaces
- Extreme involvement of new technologies
- Diversity of tools and methods
- Many Infrastructural limitations
- Telecommunication Constraints
- Incremental or Iterative methodology

- Numerous/manifold, variety of methods and tools applied

3.4 Development Team

Software Project Management is a critical task. It entails managing of different stakeholders and parties involved in development of software project. The development team plays a vital and significant role in successful project development. On the other hand, the development team can affect the success of a project in adverse direction if not handled properly. It also contributes to defining the complexity of a requirement as a team member may have his own understanding to a particular requirement because of his technical knowledge and cultural diversity. The sub-factors may be:

- Team Size
- Team Structure
- Hardworking and Focused Staff
- Diversity of Staff
- Near Shore/Offshore Team involved
- Heterogeneity of organization

3.5 Requirement Complexity Tree

The requirement complexity tree can simply be described as an analytical technique. It is a graphical representation of diverse combinations of complexity attributes. To perform measurement of complexity by using complexity tree, it is essential to indicate an undesired state of the system. This state can define the status of system's failure or even the failure of its subsystem. Then, a listing of all the possible ways, in which these events can occur, is done. Each complexity attribute is examined independently to find out the possible ways in which it can occur. Sub-trees for each requirement complexity attributes are being created independently, and the combination of all these sub-trees: requirements and expectations in Fig. 2, parties involved in Fig. 3, development characteristics in Fig. 4, and development team in Fig. 5 results in the Requirement Complexity Tree.

3.6 Requirement Complexity Tree

The sub-trees from 1–4 are the key inputs toward the development of the model. These sub-trees have been developed by analyzing the relationships among various categories of the complexity. AND and OR logic gates have been used for the same. These sub-trees contribute to construction of requirement complexity tree shown in Fig. 6.

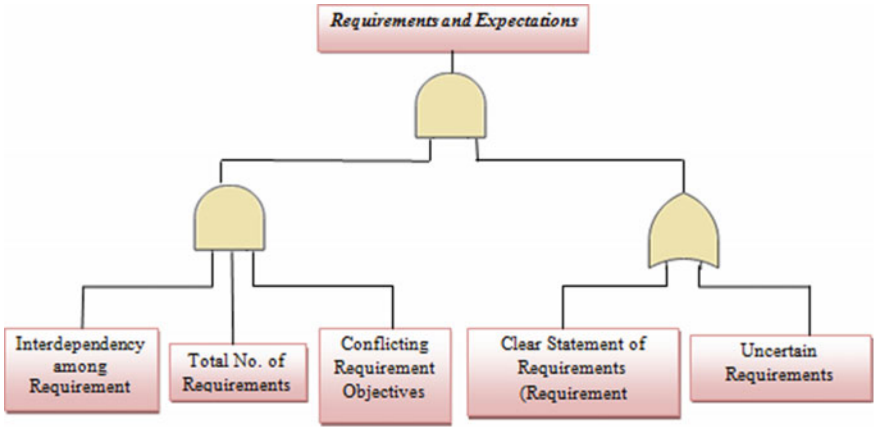


Fig. 2 Requirement complexity sub-tree (1): requirements and expectations

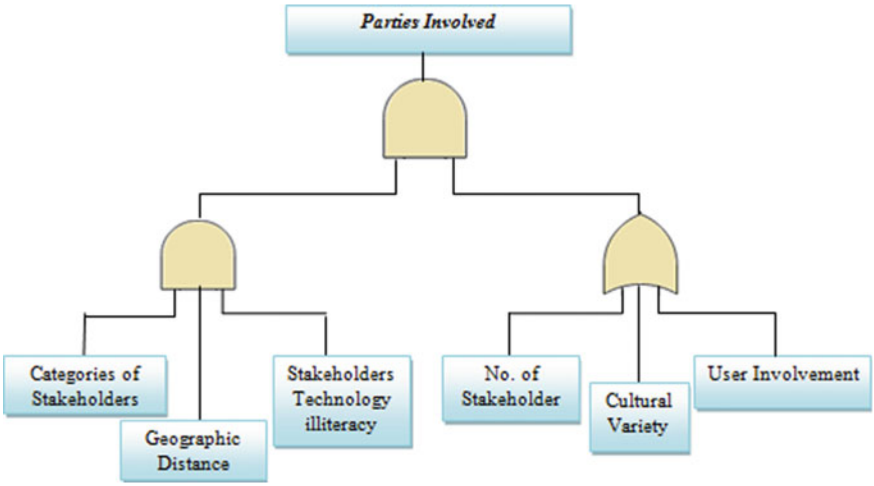


Fig. 3 Requirement complexity sub-tree (2): parties involved

4 Bayesian Network-Based Model for Requirement Complexity Prediction

After the analysis of various characteristics, Bayesian network has been finally chosen to develop a Requirement Complexity Prediction Model which works on the basis of requirement complexity attributes as discussed in section. This RCPM model uses prior knowledge and experts' opinions to predict complexity of the requirements for any software product. Formally, Bayesian network for RCPM must be trained with the data defining complexity attributes that influenced the system in the past.

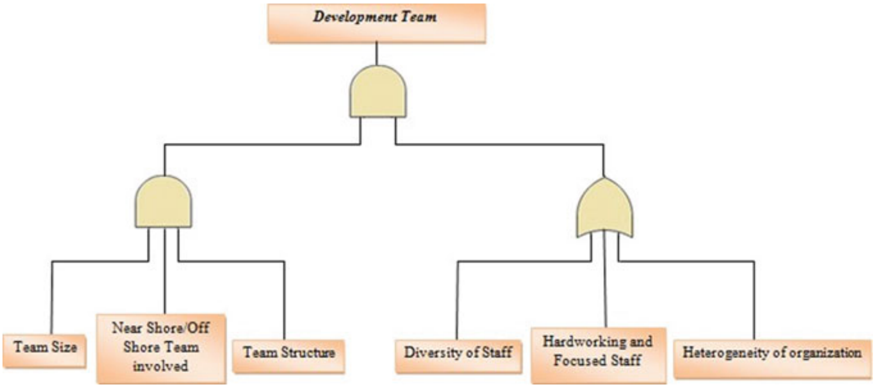


Fig. 4 Requirement complexity sub-tree (3): development characteristics

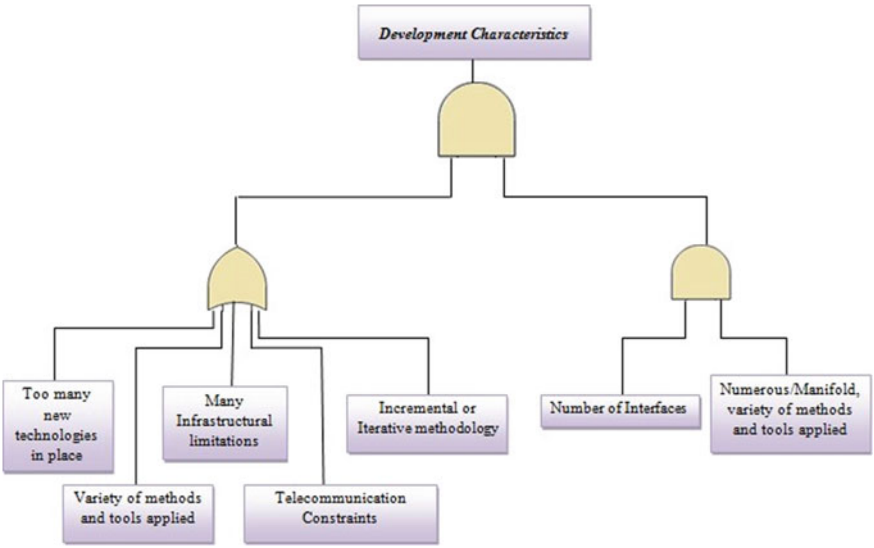


Fig. 5 Requirement complexity sub-tree (4): development team

This seeding of information is very useful to develop this Bayesian network as a useful tool. GeNIe Modeler tool has been used to construct the RCPM Bayesian network.

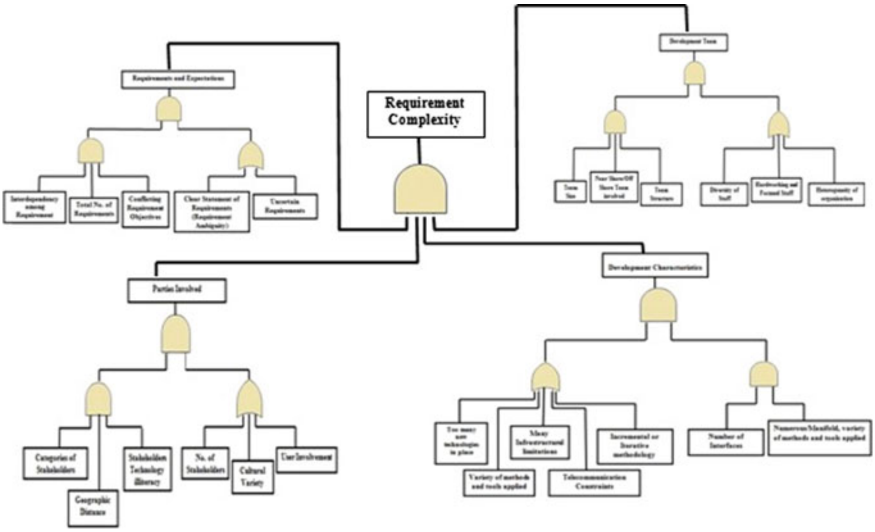


Fig. 6 Requirement complexity tree combining all sub-trees (1, 2, 3, 4)

4.1 Introduction to Bayesian Networks

Bayesian network [23] is a Probabilistic Graphical Model that is used to represent conditional dependencies between random variables. It is a directed acyclic graph where random variables are represented by the nodes and their causal influences are denoted by the directed arcs. Nodes with parents and without parents are defined through their prior probability distributions and Conditional Probability Distributions, respectively. Each and every node has a Node Probability Table (NPT) maintaining the probability distribution for every possible state of that node variable.

For instance, a node representing variable *A* can have states *A1, A2, …, Ak* and *parents(A)* is the set of parent nodes of *A*, an entry in the NPT stores the conditional probability value: *P(Ai|parents (A)*), which is the probability that variable *A* remains at state *xi* given that each of the parents of *A* is in one of its allowable state values. If *X* is a node without parents, then the NPT stores the marginal probability distributions. This allows us to represent the joint probability in the following way.

$$P(a1, a2, \ldots, an) = \Pi i = 1..n\ P(a\,i|parents(Ai)) \quad (1)$$

Once the network has been constructed, manually or using data mining tools, it constitutes an efficient mechanism for probabilistic inference. At the start, each node has its default probability values. Once the probability of a node is known because of new knowledge, its NPT is altered and inference is carried out by propagating its impact over whole of the graph; that is the probability distributions of other nodes are altered to reflect the new knowledge. This is what makes the Bayesian approach so valuable.

Characteristics of Bayesian Network

- It is generally used in uncertain situation.
- It can be used without entering any evidence by using the distributions.
- It can combine historical project data and expert opinion.
- This network is not limited to the numerical scales. It can use ordinal measures to include subjective judgments.
- It can be used for 'what-if' analysis to analyze the impact of changes in some nodes on the other nodes.

Dynamic Bayesian networks are the extensions of Bayesian network which adds a temporal dimension to the model. In other words, it can be defined as a temporal model which represents a dynamic system that changes its state usually over time [24–27]. The dynamic BN comprises a sequence of identical Bayesian networks, Zt, t = 1, 2, ..., T where each Zt represents a process that is modeled at time *t*. GeNIe Modeler from BayesFusion uses Structural Modeling, Inference, and Learning Engine (SMILE Engine) that permits interactive model building and learning.

4.2 The RCPM Model

A Bayesian network model for RCPM is constructed in GeNIe modeler. It is a dynamic Bayesian network, which combines various attributes for requirement complexity through the relationships among different attributes as defined in the requirement complexity tree. Key steps in the models are as:

Step 1. Creating sub-models for each requirement complexity attributes

Step 2. Combining the sub-models on the basis of Requirement Complexity Tree

Step 3. Defining the probabilistic distribution values

Step 4. Training the model with conditional probability table CPT (obtained from survey data)

Step 5. Testing of the model and Validation with empirical analysis

Step 1 and 2: A Bayesian model has been constructed by combining requirement complexity sub-trees (Figs. 2, 3, 4, 5).

See Fig. 7.

Step 3. Defining the probabilistic distribution values

In order to enable the model to predict the requirement complexity, the requirement analyst has to sit with his team at the early stages of the product development and must record their beliefs collectively as to the degree to which these requirement complexity factors are present in the SRS. Assuming prior information has been correctly entered into the Bayesian model; it will return a probability distribution that

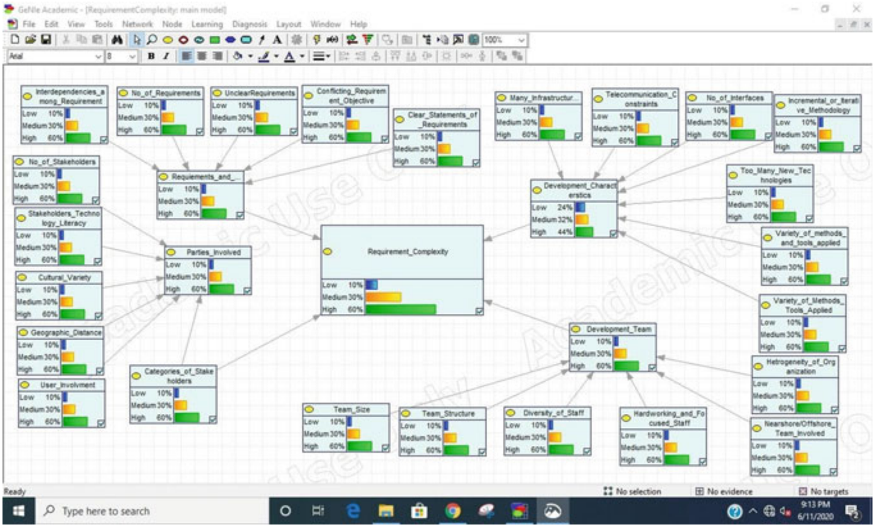


Fig. 7 Bayesian network model for requirement complexity prediction (RCPM)

can be used to predict the complexity level [28, 29]. For the BN model developed in this work, the GeNIe tool returns a probability distribution ranked between 0 and 10. Here, the attributes have been divided into 3 intervals and each interval represents a state. Rating range is shown in Table 2, in which L means Low, H means High and M means Medium. These beliefs are required to be entered into Bayesian network on the basis of these scales and intervals.

Similar probability range and values has been used for the different attributes of requirement complexity components.

Table 2 Probability distribution range

Requirement complexity component	State	Intervals
Requirements and expectations	*H*-High	7–10
	M-Medium	3–7
	L-Low	0–3
Parties involved	*H*-High	7–10
	M-Medium	3–7
	L-Low	0–3
Development characteristics	*H*-High	7–10
	M-Medium	3–7
	L-Low	0–3
Development team	*H*-High	7–10
	M-Medium	3–7
	L-Low	0–3

Step 4. Training the model with Conditional Probability Table (CPTs)

The proposed model has been seeded with the requirement data. The relationship between a child node and its parent is entered by an equation. This is possible in all the cases: Nodes are discrete or continuous and the relationship is deterministic or probabilistic [30–35]. GeNIe converts all equations into tables prior to compilation of a network. The following inbuilt function is used as an equation to build the CPTs of the dependent nodes. This distribution has a graph of Triangular shape, with nonzero values from a−w to a + w and the highest point at x = a.

Triangular dist function:	
Usage	TriangularDist (x, m, w)
Definition	$(w-\|x-m\|)/w^2$
Required	$w > 0$
Support	$M-w \leq x \leq m + w$
Moments	$\mu = m\ \sigma = w/\sqrt{6}$ $\gamma 1 = 0\ \beta 2 = 2.4$

Equations of all the dependent nodes present in the network are like:

p (**RequirementandExpectations**|Total number of requirements, Interdependency among requirement, Conflicting Requirement Objectives, Clear Statement of Requirements, Uncertain Requirements, Changing and Regulatory Requirements) = TriangularDist (RequirementandExpectations, Total number of requirements, Interdependency among requirement, Conflicting Requirement Objectives, Clear Statement of Requirements, Uncertain Requirements, Changing and Regulatory Requirements)

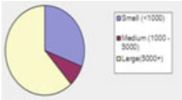
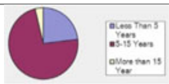
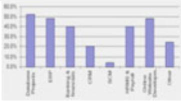
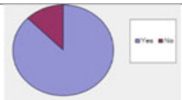
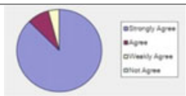
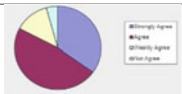
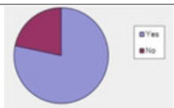
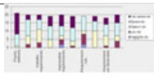
Step 5. Testing of the model and Validation with empirical analysis

The proposed requirement complexity attributes to measure the requirement complexity were taken into consideration through the extensive literature review. The proposed measurement tool has been seeded with the beliefs of the practitioners and experts working in this specific domain collected through a structured survey. The model has been validated by performing a case study on some real-time project. These experts and practitioners involved in the survey were contacted through social networks and personal contact.

The objective of this survey was to collect information about the complexity attributes and use experts' belief to identify major complexity attributes that affect a successful development of software projects.

The result of the survey (Parts 1 and 2) has been summarized in Table 3, and results for Part 3 have been used in developing the Bayesian network.

Table 3 Summary of the survey results (Parts 1 and 2)

Survey questions	Pie chart of the results	Analysis
Company size of the respondents		Small (<1000): 30.8% Medium (1000–5000): 7.7% Large (5000 +): 61.5%
Work experience		Less than 5 year: 23.1%, 5–15 year: 73.1%, More than 15 years: 3.8%
Type of developed projects		Database: 52.0%, ERP: 48.0%, Financials: 40.0%, CRM: 20.0%, SCM: 4.0%, HRMS 40.0%, Web development: 48.0%, Other: 24.0%
Requirement complexity can be considered as a risk to software development		Yes: 87.5% No: 12.5%
Necessity to identify and predict requirement complexity at early stage of development		Strongly agree: 87.5%, Agree: 8.3%, Weakly agree: 4.2%, Not agree: 0.0%
Responses regarding requirement complexity components		Strongly agree: 34.8%, Agree: 47.8% Weakly agree:13.0%, Not agree: 4.3%
Responses for requirement complexity attributes can be considered as interlinked factors		Yes: 83.17% No: 16.83%
Responses for ranking the various requirement complexity attributes according to their severity		Components and sub-trees got the rankings that have been used to build the Bayesian network

5 Validation of the RCPM model

In order to validate the proposed model, an academic MIS project "SMIS" has been chosen as summarized below. Table 4 gives a brief description of the project.

Bayesian network generates Adjacency Matrix for the components and attributes of the requirement complexity. This matrix has been used to validate the probability distribution and RCPM model (Fig. 8).

Table 4 Description of SMIS project

Team size	4 members
Scheduled time	6 months
Project category	Web-based project
Platform	WINDOWS
Technology	C# and SQL server
No. of functional requirements	80

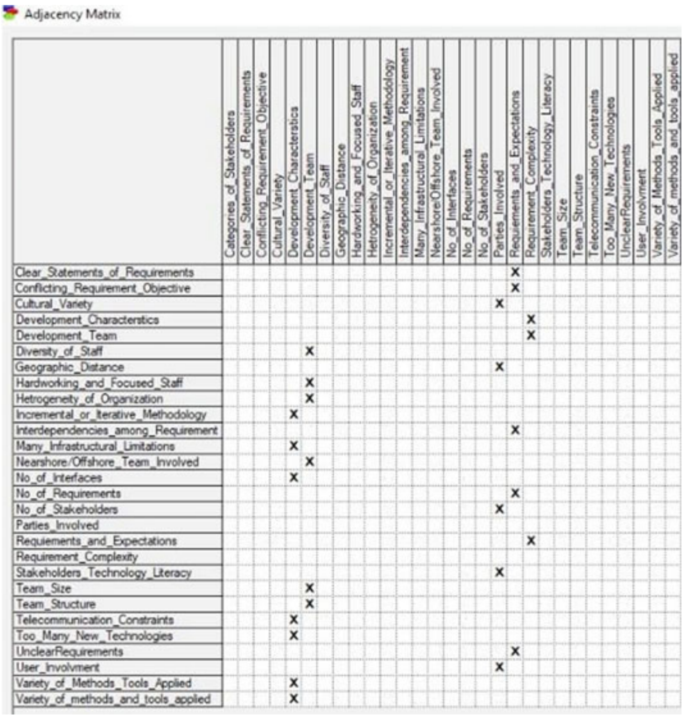


Fig. 8 Adjacency matrix for requirement complexity components

RCPM based on Bayesian network has been used to analyze the requirement complexity risks. Firstly, a formal inspection was performed by the development team to look for the potential requirement complexities and have been labeled from RC1 to RCN. Then, the survey data has been entered into the Bayesian model. A total of 80 requirements were in the project that has been analyzed for identification of complexity attributes, and results have been summarized in Table 5.

Table 6 is the report generated after compiling the BN for Severity. After compiling the BN, select the abovementioned node values and their respective severity factors. This is what you get as a result.

Table 5 Requirement complexity attributes

Requirements	Requirement complexity attributes			
	RC1	RC2	RC3	RC4
R1	N	N	N	N
R2	Y	N	N	N
R3	N	N	N	y
R4	Y	N	N	N
R5	N	N	N	N
R6	N	N	N	N
R7	N	N	N	y
R8	N	Y	Y	N
R9	Y	N	N	N
R10	Y	N	N	N
TOTAL	RC1 = 4	RC2 = 1	RC3 = 1	RC4 = 2

As per the RCPM, the predicted value for Requirement Complexity is lying in medium range. This predicted value can be shared with the stakeholders, and a negotiation can be done on the agreed set of requirements to minimize the complexity.

6 Conclusion

Complexity in software requirement specification generally results in high development cost and delays in schedule as to rework on software requirement. Even it sometimes causes software failures.

This work proposes a Bayesian network-based model to predict requirement complexity at an early stage of the software development. The Requirement Complexity Prediction Model works on the basis of four complexity attributes identified through extensive literature review.

The proposed model has been implemented in GeNIe modeler. The model has been seeded with experts' beliefs about related to different complexity attributes. This prediction model has been validated through a case study. The result shows that the proposed model is capable of predicting the complexity at an early stage. These results can be useful for a requirement analyst as to discuss this predicted complexity values with his team which can be further discussed with the stakeholders to negotiate the requirements if required. The Future Work involves working on a larger dataset to test the efficiency of the model.

Table 6 Custom report generated from GeNIe

Requirement complexity attributes	Status		
	Low	Medium	High
Total no. of requirements	0.024741	0.4146	0.56066
Interdependency among requirement	0.40917	0.46836	0.12246
Conflicting requirement objectives	0.20904	0.34277	0.44819
Clear statement of requirements	0.14493	0.31339	0.54168
Uncertain requirements	0	1	0
No of stakeholder	0	0	1
Categories of stakeholders	0.038852	0.39111	0.57004
Cultural variety	0.075107	0.26895	0.65594
Geographic distance	0	1	0
Stakeholder technology illiteracy	0.078475	0.30365	0.61787
User involvement	0.18938	0.45958	0.35103
Number of interfaces	0.98025	0.01017	0.0095838
Too many new technologies in place	0.32894	0.46874	0.20232
Team size	0.1378	0.26535	0.37769
Team structure	0.072634	0.069862	0.18929
Hardworking and focused staff	0.020953	0.031742	0.13928
Diversity of staff	0	0	1
Near shore/offshore team involved	0.23211	0.33208	0.16911
Heterogeneity of organization	0	0	0
Numerous/manifold, variety of methods and tools applied	0.12928	0.42288	0.14891
Incremental or iterative methodology	0.53715	0.30399	0.056094
Telecommunication constraints	0	0	0
Many infrastructural limitations	0.10009	0.053458	0.34357
Variety of methods and tools applied	0.004717	0.014567	0.090932
Requirements and expectations	0.34357	0.69325	0.25678
Parties involved	0.124587	0.68785	0.4523
Development characteristics	0.2356	0.4869	0.5478
Development team	0.4698	0.3245	0.3578
Requirement complexity	–	0.15789	–

References

1. Christel MG, Kang KC (1992) Issues in requirements elicitation, technical report 1992 Software Engineering Institute, Carnegie Mellon University Technical Report 1992. Available at https://resources.sei.cmu.edu/asset_files/TechnicalReport/1992_005_001_16478.pdf
2. Charette RN (1989) Software engineering risk analysis and management. Intertext Publications, New York, NY, USA

3. Sakthivel S (2007) Managing risk in offshore systems development. Commun ACM 50(4):69–75
4. Schmidt R, Lyytinen K, Keil M, Cule P (2001) Identifying software project risks: An international Delphi study. J Manag Inf Syst 17(4):5–36
5. Tafti MHA (2013) Risks factors associated with offshore IT outsourcing. Ind Manag Data Syst 105(5):549–560
6. Dennis MA (ed) (2015) The Chaos Report. Standish Group International, Inc
7. Murray JP (2000) Reducing IT project complexity. Inf Strategy: Executive's J 16(3):30–38
8. Lyytinen K, Hirschheim R and R. Hirschheim Paper (1987) Lyytinen K and Hirschheim R Paper 1987.pdf. Oxford Surveys in Information Technology 4:257–309
9. Ewusi-Mensah K (2003) Software development failures anatomy of abandoned projects
10. Zuse H (1992) Properties of software measures. Softw Qual J 1:225–260. https://doi.org/10.1007/BF01885772
11. Banker R, Datar S, Zweig D (1989) Software complexity and maintainability. 247–25:5. https://doi.org/10.1145/75034.75056
12. Kushwaha DS, Misra AK (2006) A complexity measure based on information contained in software. In: 2006 Proceedings of the 5th WSEAS international conference on software engineering, parallel and distributed systems, Madrid, Spain, February 15–17, pp 187–195
13. Sadia H, Abbas SQ, Faisal M (2019) A systematic literature review of multi-criteria risk factors (VUCA) in requirement engineering. Int J Sci Technol Res 8(11). ISSN 2277–8616 13
14. Sharma A, Kushwaha DS (2010) A complexity measure based on requirements engineering document. J Comput Sci Eng 1(1)
15. Sharma A, Kushwaha DS (2012) Applying requirement based complexity for the estimation of software development and testing effort. ACM SIGSOFT Softw Eng Notes 37(1)
16. Purawinata WM (2017). The prediction of software complexity based on complexity requirement using artificial neural network ieee international conference cybernetics and computational intelligence
17. Keshavarz G, Modiri N, Pedram M (2011) Metric for early measurement of software complexity. Int J Comput Sci Eng (IJCSE) 3(6)
18. Mas JAD, Medina DL (2018) Measuring the project management complexity: the case of information technology projects. Complexity Hindawi, https://doi.org/10.1155/2018/6058480
19. Chapman RJ (2016) A framework for examining the dimensions and characteristics of complexity inherent within rail megaprojects. Int J Project Manage 34(6):937–956
20. Vidal L, Marle F (2008) Understanding project complexity: implications on project management. Kybernetes 37(8):1094–1110
21. Ireland V, Rapaport B, Omarova A (2012) Addressing wicked problems in a range of project types. Procedia Comput. Sci 12:49–55
22. Owens D, Ahn J, Shane J, Strong J, Gransberg KC (2012) Defining complex project management of large U.S. Trans Proj 17(2):170–188
23. Xia B, Chan APC (2012) Measuring complexity for building projects: a Delphi study. Eng Constr Archit Manag 19(1):7–24
24. Jensen F (2001). Bayesian network analysis tools. https://doi.org/10.1007/978-1-4757-3502-4_6
25. Lee P (1989) Bayesian statistics, an introduction, 2nd edn. Oxford university Press, New York
26. Von Winterfeldt D, Edwards W (1986) Decision analysis and behavioral research, Cambridge University Press MA
27. Laskey K (2003) Bayesian inference and decision theory. SYST 644 Class Notes Department of Systems Engineering. George Mason University. http://ite.gmu.edu/-klaskey/SYST664/SYST664.html Spring 2003
28. Wooff DA, Goldstein M, Coolen FPA (2002) Bayesian graphical models for software testing. IEEE Trans Software Eng 28(5):510–525
29. Ashby D (2006) Bayesian statistics in medicine, Paper Celebrating the 25th Anniversary of Statistics in Medicine. STATISTICS IN MEDICINE

30. Stamelos I, Angelis L, Sakellaris E (2003) On the use of Bayesian Belief network for the prediction of software productivity. Inf Softw Technol
31. Ward S, Chapman (2003) Transforming project risk management into project uncertainty management. Int J Proj Manage
32. Fenton N, Neil M Predicting software defects in varying development lifecycles using Bayesian nets. Inf Softw Technol 49
33. Wooff AD, Goldstein M, Coolen PAF (2002) Bayesian graphical models for software testing. IEEE Trans Softw Eng
34. Wu YP, Hu QP, Poh KL, Ng SH, Xie M (2005) Bayesian networks modeling for software inspection effectiveness. In: I th pacific rim international symposium on dependable computing
35. Cooper G (1990) Computational complexity of probabilistic inference using Bayesian belief networks. Artificial Intelligence

Anomaly Detection Using Feature Selection and Ensemble of Machine Learning Models

B. Anbarasu and I. Sumaiya Thaseen

Abstract Vulnerabilities have increased in cyberspace. This is due to the technology growth and the huge amount of data communicated between various endpoints of a network. Intrusion Detection System (IDS) plays a major role in identifying malicious traffic in the network. Intrusion Detection is much challenging because of the raw network traffic it contains with a large number of attributes that add to the complexity of the model. Several Machine Learning (ML) models have been built to solve these issues. The problem even after the new technologies' introduction is the lack of datasets, classifiers work best for one problem and serve the least for the other set of problems, decision algorithms are not framed effectively. To overcome these problems, proposed method uses an ensemble approach on classifiers to provide a better solution for feature selection parameters. Our model outperforms the accuracy and detection rates compared to the individual classifiers. Decision Tree (DT), Logistic Regression (LR), and Support Vector Machine (SVM) values are given to the Random Forest (RF) ensemble classifier. Experiments are performed using CICIDS 2017 dataset. The proposed approach has accuracy of 98%, recall 97%, precision 100%, and F-score 98% compared to the individual models.

Keywords Ensemble classifiers · Intrusion detection system · Machine learning · Network traffic · Optimizer

1 Introduction

1.1 Cyber Attacks

The Internet has both positive and negative effect on the end users. The computation in the cyberspace is growing exponentially with the needs of the consumers. A device

B. Anbarasu (✉) · I. S. Thaseen
School of Information Technology and Engineering, Vellore Institute of Technology, Tamil Nadu, Vellore, India
e-mail: araskrishna@gmail.com

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_16

is susceptible to attack once it is connected to the Internet. Attacks are growing at a faster rate, and hence, there is a need to develop a secure and robust system. Several types of attacks are Brute force attacks, SQL injection, Zero-day attacks, Denial-of-Service (DoS), Distributed Denial-of-Service (DDoS), Domain Name System (DNS) Tunneling, and Man-in-the-Middle attacks that target the cyberspace. The knowledge needed to attack cyberspace is much easier because various tools are readily available in the market. Attack detection has to be the highest priority [1].

1.2 Anomaly Detection

Anything that deviates from a regular process is termed Anomaly detection or Outliers. Anomaly detection occurs in a variety of applications like medicine, financial transactions, network traffic, Fraud detection in credit card transactions, detecting fake news, industry applications, etc., and in other similar events which are not easier to detect. Challenges of Anomaly detection include feature learning, defining behavior to be normal or abnormal, and distribution of data, spatiotemporal behaviors are some among others [2]. Anomaly detection is not easier as it occurs very rare one in a million or billion. It is a negotiable metric, but the harm it may create is unexpected. Anomaly detection is determined using proximity-based, density-based, and cluster-based approaches. Such techniques are not generic; a solution for a particular problem may not be the solution for the other. In certain cases, the noise is also misclassified as an anomaly; the greater concern is the lack of real-time datasets. Data classified to be normal may not be normal forever. ML serves as the best approaches in detecting anomalies. Anomaly detection models are capable of identifying previously unknown attacks [3]. Some of the anomaly detection techniques are classification, clustering, and statistical approaches [1].

The contributions of this study are as follows:

- Developing an ensemble anomaly detection using ML models.
- Hyper-parameter Tuning is applied in the ensemble model for optimization.
- Ensemble model results in better performance metrics when compared to the individual models.

The rest of this paper is as follows: Section 2 showcases the relevant researches for anomaly detection in ML models; The background of ML models, data preprocessing, and feature selection is discussed in Sect. 3; Sect. 4 outlines the proposed design with data parameter optimization and ensemble classifiers; Sect. 4 provides results of the problem with their performance metrics followed by conclusion at Sect. 5.

2 Literature Review

This section briefs about anomaly detection using ML models. Artificial Intelligence (AI) framework constitutes monitoring agent, AI-based reaction agent. ML models analyze the network patterns to classify the anomalies in IDS in IoT systems [4]. The learning is performed by observing the multiple event sequences and some events are removed and analyzes the frequently observed sequences. However, detection of an anomaly for single, rarely occurring events depends only on the operation condition. The evaluation is done to only specific conditions that also suffer from insufficient data [5]. A semi-supervised learning algorithm for collaborative IDS is used to mitigate the attacks whereas traditional models suffer from instances of labeling. The limitation is heavy cost and vulnerability to insider attacks [6]. Classical ML problems with quantum computing make it easier in learning tasks. But quantifying the computing is very challenging for applications. Complex data provides a simple and rigorous learning problem for fault-tolerance [7]. Different ML techniques are applied on the KDD dataset. The dataset is pre-processed with J48, Multi-Layer Perceptron (MLP), and Bayes Network classifiers. High accuracy is obtained for the J48 classifier and classifies all types of attacks DOS, R2L, U2R, and PROBE. However, more classifiers have to be tested for feature selection to exploit the most important features [8].

Among the various tree-based classification algorithms, RF has the best performance metrics compared to the other counterparts. NSL-KDD 99 dataset is deployed for the experiments. Accuracy of this model is 97.49% [9]. ML integrated with feature selection techniques plays an important role in IDS. A combination of feature selection techniques has better results for NSL-KDD 99 dataset. K-Nearest Neighbor (k-NN) classifier with Information Gain Ratio (IGR) feature selection has an accuracy of 99.07% [10]. Processed data plays an important part in the ML/Data mining (DM) approach for IDS [11]. IDS effectiveness is decided based on the dimensionality of the data and removing the irrelevant attributes, thereby reducing the computational complexity for the NSL-KDD dataset. The performance of RF, Extra Tree Classifier (ETC), and DT is above 99% for attack classes [12]. Classification techniques must be effective and quick in classifying benign and malicious activities. Different classifiers tested on NSL_KDD dataset are J48, RF, and Naïve Bayes (NB). Performance is greater than 98% for J48, and RF whereas in the case of NB it is 81% [13].

Nearly 12 ML algorithms are tested for the smart grid systems with CICIDS 2017, UNSW-NB15, and the Industrial Control System (ICS) cyberattack datasets. The RF algorithm has better performance with 99.9% in comparison with the other algorithms. The selection of ML techniques is based on the application system [14]. The model performance gets affected by the number of anomalies in the real traffic data. Class imbalances affect the performance of SVM and, RF when the attack proportion changes dynamically over time [15]. Anomaly detection for ML based on the spatial–temporal features and methods for packet sequences are tested [16]. The classifier RF-800 outperforms other models while using NSL-KDD dataset in performance but using UNSW-NB15 dataset accuracy is better for DT [17]. Four feature selection algorithms along with two ML algorithms are used to determine the performance using CICIDS 2017 dataset. In the initial phase, features are selected using the respective algorithms and then implemented using ML, and results are analyzed [18]. When the number of features increases, the computational complexity also increases. This can be reduced using information gain which selects the relevant and significant features from the CICIDS 2017 dataset. The algorithms RF and J48 produce an accuracy of 99.86% and 99.87%, respectively; but, in J48 the execution time is more [19].

3 Background

In this section, various techniques used in the data pre-processing, feature selection, dimensionality reduction, and classifiers are analyzed.

3.1 Data Pre-processing

Raw data has to be processed to make it suitable for the ML model. It is an integral part of ML because the results of the classifier have a negative impact if the pre-processing is not done. So pre-processing of data is more important to the model. The dataset contains redundant and irrelevant data resulting in a class imbalance problem. Real-world datasets contain noise, missing information that has to be pre-processed before it is sent to the proposed model. Therefore, a data transformation is necessary. If there are more null values or missing values in a particular row or column, the same can be deleted, such that the class imbalance problem is solved. The best option to determine missing values is to calculate mean/mode/median and to replace the missing value with that particular value. This will solve the data loss problem, and it serves better than deleting the row/columns. ML algorithms work on mathematical equations, so it is not possible to work with categorical data so that values have to be encoded. Hence, one hot encoding, label encoder can be used for this purpose. Label encoder categorizes the values with a default number for similar values which results in a conflict as the numbers may be given priorities. Therefore, one hot encoding is used to overcome the limitation. It creates additional columns from label encoding and checks the existence of values to fill with 0 or 1. Then, this dataset is divided into train set and test set. The test set is used for predicting the results, the usual combination in 80:20 or 70:30 rules to split the data. Feature scaling is used to standardize the values in a fixed range to avoid the dominance of one column to other resulting in wrong results. Feature scaling can be done using min–max normalization and standardization [20, 21].

(i) **Min–Max Normalization**

It is the most common way to normalize the data and preserves the relationship among the existing values. It is the linear transformation to the original data. The method with the transformation of value 0 for minimum value and 1 for the maximum value and the remaining values are transformed in between 0 and 1. The representation $\min_A$ and $\max_A$ are minimum and maximum values of an attribute A which maps to the value v_i [22, 23].

$$v_i' = \frac{v_i - \min_A}{\max_A - \min_A}(\text{new\_max}_A - \text{new\_min}_A) + \text{new\_min}_A \tag{1}$$

(ii) **Standardization**

In this method, the values are normalized based on the mean ($\overline{A}$) and the standard deviation (σ_A) is also called as Z score normalization v_i the value, and v'_i mapped value. Here, the mean values are zero and the variance equals the value of 1 [23].

$$v_i' = \frac{v_i - \overline{A}}{\sigma_A} \tag{2}$$

3.2 *Feature Selection*

Feature selection reduces the computation cost and improves the performance of the models. The number of input variables has to be reduced to reduce the complexity of the system. Chi-square metric is used along with classification strategies for optimal results.

Chi-Square

It measures the deviation for the feature event independent of the class value from the expected distribution [24]. The metrics are true positives (t_p), true negatives (t_n), false positives (f_p), false negatives (f_n), probability of the number of positive cases P_{pos}, and negative cases P_{neg}

$$\begin{aligned}\text{chi\_ square\_ metric} = t(t_p, (tp + fp)P\text{pos}) + t(f_n, (f_n + t_n)P_{pos}) + \\ t(f_p, (f_p + t_p)P_{reg}) + t(t_n, (t_n + f_n)P_{reg})\end{aligned} \tag{3}$$

Here,

$$t(\text{count}, \text{expect}) = (\text{count} - \text{expect})^2/\text{expect} \tag{4}$$

3.3 Decision Tree

The Decision Tree (DT) is a supervised learning classifier with low computational complexity. The rules are configured in a very simple manner to understand resulting in better detection accuracy for known intrusions. A tree structure has several nodes with multiple output branches. It effectively identifies benign and malicious traffic based on the predefined parameter for splitting the data into training and testing. This predictive ML model maps between the values and attributes of objects. Commonly used decision tree models are ID3, CART (Classification and Regression Trees), and Reduction in Variance.

3.4 Random Forest

A Random Forest (RF) is a simple, robust ML ensemble technique used to solve problems in classification and regression. It is proposed by Breiman in 2001 [25] with minimum fine-tuning on the best features which can be used for the model. The accuracy of RF is equal to AdaBoost and sometimes outperforms it. It is faster than bagging and boosting [26]. Many classifiers are integrated to solve complex problems such as regression and classification problems. Many decision trees are combined, and the ensemble learning technique integrates many classifiers to provide solutions for more complex problems. DT and RF use the same hyper-parameters as DT, but RF outperforms DT. The classifier RF works effectively on high-dimensional data with less computational complexity.

3.5 Support Vector Machine (SVM)

SVM works effectively on small datasets and also works well with high-dimensional spaces. It supports both classifications as well as regression. It can easily ignore the outliers. Complex data transformation is carried out using SVM, and later, it separates the data based on the outputs. In addition, parameters can also be tuned to get better results. This model results in higher classification accuracy due to the amount of training time spent and aids in dimensionality reduction. SVM classifier tries to reduce the generalization error by increasing the training data margins as given in Eqs. (5) and (6) [29].

Thus,

$$\min_{w,b} \frac{1}{2} \|w\|^{-2} \tag{5}$$

Such that

$$y_i(w^T x_i + b) \geq 1, i = 1, 2, \ldots\ldots, M \tag{6}$$

where 'w' represents n-dimensional vector, b is a scalar constant, x_i specifies the data point, and y_i is class of positive value with $y_i = 1$ and negative value with $y_i = -1$.

3.6 Logistic Regression (LR)

LR is a supervised learning classifier for calculating the target variable. It briefs about the association of dependent and independent variables [12]. The algorithm works better for the same input parameters compared to the previous instance as more relevant data is fed into the model. It predicts categorical class labels and classifies the data and models to predict unknown or missing values. LR models the data using the sigmoid function is given in Eq. (7) [29].

$$\mathrm{g(z)} = \frac{1}{1 + e^{-z}} \tag{7}$$

where 'z' represents the sample, g(z) is sigmoid function, and e is the Euler's number.

3.7 Ensemble Design Model

The basic idea of voting is the linear combination of learners is the easiest way to combine multiple classifiers [30] where

$$y_i = \sum_j w_j d_{ji} \tag{8}$$

where

$$w_j \geq 0, \sum_j w_j = 1 \tag{9}$$

d_{ji} are base learners with values between −1 and 1.

The fundamental concept of majority voting is that each classifier initializes the votes. The final result will be the opinion with the highest votes. Littlestone and Warmuth [31] stated that by introducing weights into the majority voting technique, the number of mistakes in the ensemble model can be reduced. The median rule used for outliers is more robust in ensemble models.

4 Proposed Design

In this section, the proposed anomaly detection model is detailed. Figure 1 shows the proposed anomaly detection model. The architecture constitutes data cleaning, feature selection, and core design algorithm. A conversion process is implemented to transform the extracted features into numerical form. The data is sampled using under-sampling. Feature selection is implemented using chi-square approach. Classifiers receive the processed dataset for decision-making to decide whether it is normal or abnormal. The decision is updated in the database to check for the next values, and this process continues for all the values. The framework is robust and agile in detecting the anomalies in IDS.

4.1 *Automatic Parameter Optimization*

The results of feature selection are sent to three classifiers namely DT, LR, and SVM. The results are integrated using Random Forest Classifier. CICIDS 2017 dataset is used for the experiments.

In the proposed model, parametric optimization is implemented. Single objective optimization may not provide the best performance criteria but multi-objective optimization would perform better by identifying the optimal conditions of the constraints defined for the system. Model performance is decided based on the choice of the classifier. These parameters are associated with the classifiers which are used in the model. The complexity of the model and best optimal performance should have a

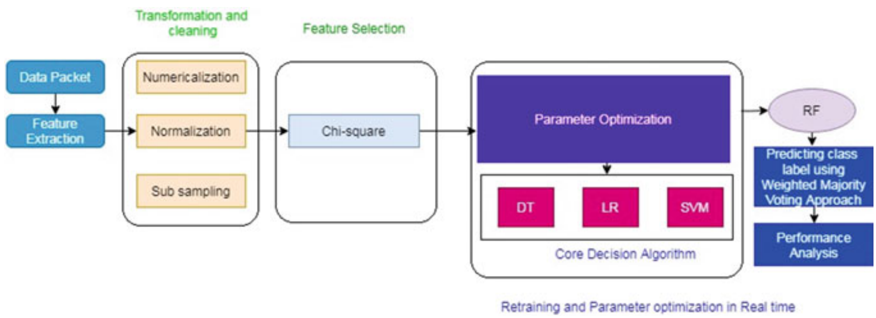


Fig. 1 Proposed framework of anomaly detection IDS

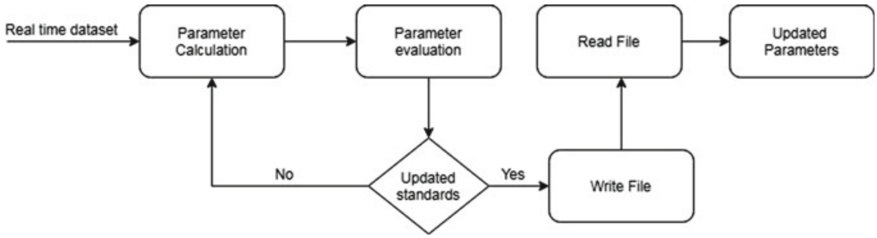


Fig. 2 Auto-para optimization model

strong correlation to achieve a high detection rate in anomaly prediction. Figure 2 shows the parametric optimization approach.

4.2 *Ensemble Classifiers*

Ensemble produces higher accuracy and detection rate compared to the single classifier. Single classifiers cannot provide better accuracy because of their individual limitations. The proposed model deploys DT, LR, and SVM classifiers, and these outputs are sent to RF. These combined classifiers will result in better performance. Several models output is integrated to get an efficient classifier ensemble. The weighted majority voting approach is used for merging the predictions. Each of the classifier predicts a different class label, and the majority class label is the final result of the approach. Classifiers categorize the data based on the required number of classes and subclasses. The ensemble method produces better predictions and improved performance. The model fitting is performed several times on training datasets, and the predictions are combined resulting in accurate mean, mode for regression, and classification, respectively. These ensemble classifiers provide the best accuracy, better variance, and bias.

5 Results

5.1 *Dataset Description*

CICIDS 2017 dataset is used for experimental analysis. This dataset contains the most updated common attacks of real-world data. The network traffic analysis is captured using CICFlowMeter. The dataset has realistic background traffic built using complete network configuration and capturing traffic analysis with various attacks in heterogeneous environments. Table 1 lists out the number of samples available in the dataset along with its category.

Table 1 Different category samples of CICIDS 2017 dataset

Category	Samples
Benign	168,186
Brute force	1507
XSS	652
SQL injection	21

5.2 Experimental Setup

The experiments are executed on the Google Colab platform using 'engine = python' to avoid the UnicodeDecodeError: 'utf-8' encoding error. The dataset file contains 458,968 records and 85 columns. Certain columns have redundant values, so those columns are removed from the file such as FwdHeader Length and FwdHeader Length 1. These columns have identical value, the latter is being removed, and the file is updated with 84 columns. Several blank records exist in the file and some file contains the non-numeric value which has been replaced with the respective data types. Infinity values and NaN values are replaced with − 1 and to replace the string with numeric we use label encoding.

For the experiments, total number of benign records is 168,186 and the attack record is 2180 which accounts to 1% of the total records. Thus, a data imbalance is present. It can be solved using the under-sampling technique. Features with ports and addresses are susceptible to attacks. Therefore, these values can be substituted during training and those columns are removed. A manual under-sampling is performed. The entire attack samples are utilized, and the remaining records are copied from the benign records of 70% which is 5087 records. This new benign record is copied from the total benign record based on the probability not exceeding the count of 5087 as shown in Table 2.

For the benign records, the value is encoded as 0 and the attack record is encoded as 1. Nearly 7 columns are removed from the dataset because these attributes do not contribute to the performance of the model. In addition, these values can be used by the intruder so these are removed [27, 28]. The seven column names are as follows: Flow ID, Source IP, Source Port, Destination IP, Destination Port, Protocol, and Timestamp.

Figure 3 plots the feature importance graph of the CICIDS 2017 dataset. The correlated features are removed from the dataset, and hyper-parameter optimization is

Table 2 Balanced category samples of the CICIDS 2017 dataset

Category	Samples
Benign	5087
Brute force	1507
XSS	652
SQL injection	21

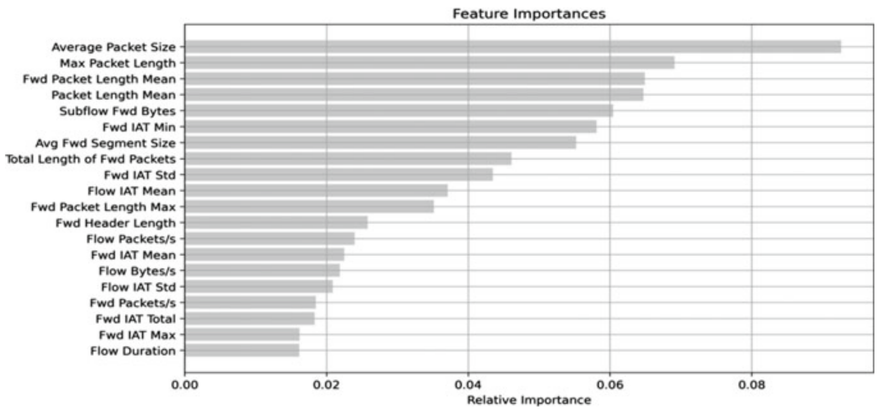


Fig. 3 Feature importance graph on CICIDS 2017 dataset

Table 3 Features and importance of classifier

S. No.	Feature number	Importance	Features
1	#51	0.081	Average packet size
2	#13	0.076	Flow bytes/s
3	#38	0.064	Max packet length
4	#7	0.055	Fwd packet length mean
5	#23	0.054	Fwd IAT min
6	#3	0.039	Total length of Fwd packets
7	#15	0.036	Flow IAT mean
8	#5	0.032	Fwd packet length max
9	#21	0.029	Fwd IAT Std
10	#33	0.027	Fwd header length

applied for the model. In the proposed model, grid search along with cross-validation is applied for better performance (Table 3).

5.3 Evaluation Parameters

The various metrics used for analyzing the performance are accuracy, precision, recall, and F-score. Accuracy predicts the performance of the system which is calculated using the ratio of perfectly predicted values to the total number of values. Accuracy and performance are directly proportional. Precision is the ratio of correctly predicted observations to the total number of correct observations. An increase in

Table 4 Comparison between ML models

Metrics	DT	LR	SVM	Proposed
Accuracy	72.22	83.99	94.41	98.44
Precision	67.26	84.46	94.32	100
Recall	86.53	83.30	94.51	96.88
F-score	75.68	83.88	94.41	98.42

precision reduces the false positive. Recall determines the correctly predicted observations. The evaluation metrics are listed below in Eqs. (8), (9), (10), and (11), where True Negative (TN), True Positive (TP), False Negative (FN), and False Positive (FP) [29].

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{FN} + \text{FP} + \text{TP}} \tag{10}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{11}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{12}$$

$$\text{F} - \text{score} = 2{*}\frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \tag{13}$$

Table 4 specifies the comparison of ML models by ensemble the three individual classifiers proposed in the model. Hyper-parameter optimization using grid search with cross-validation produces best results with the RF classifier integration. The proposed model results are better in comparison with other traditional methods. Its accuracy is better than the individual classifiers.

Figure 4 plots the Receiver Operating Characteristics (ROC) Curve of the CICIDS 2017 dataset for different classification models used in the proposed approach with the parameters TPR and FPR. Hyper-parameter tuning is applied to improve the performance measures such as F-score and accuracy values of training and testing data. The proposed model automatically selects the parameters using the grid search algorithm. Grid search cross-validation results are plotted in Fig. 5.

6 Conclusion

Cyberattacks are evolving every day and have become a threat to modern systems. There is a need to secure network traffic by analyzing the malicious behavior. In this paper, an ensemble approach is developed for anomaly detection. The proposed

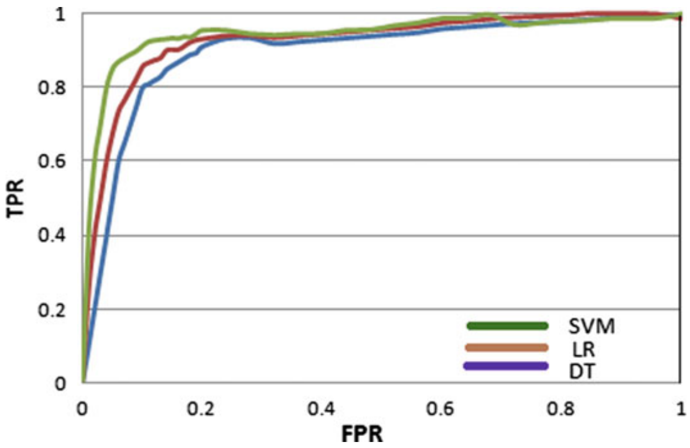


Fig. 4 ROC on CICIDS dataset

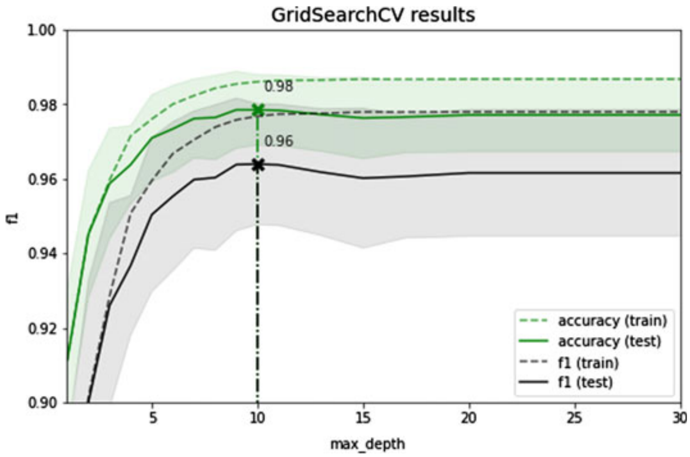


Fig. 5 Grid search CV results

model is built using chi-square feature selection and ensemble of supervised classifiers. The ensemble approach integrates DL, SVM, and LR, and the results are sent to RF classifier. The experiments are performed on CICIDS 2017 dataset. The proposed model has better feature selection results, and the performance results are better in comparison with the individual classifiers. Automatic parametric optimization of the individual classifiers is deployed using grid search. The proposed approach has an accuracy of 98%, recall of 97%, precision of 100%, and F-score of 98% compared to the individual ML models. Our future work will concentrate on implementing real-time network data in deep learning models for recently available datasets and fine-tune the models for better performance.

References

1. Ahmed M, Mahmood AN, Hu J (2016) A survey of network anomaly detection techniques. J Netw Comput Appl 60:19–31
2. Biradar KM, Gupta A, Mandal M, Vipparthi SK (2019) Challenges in time-stamp aware anomaly detection in traffic videos. arXiv preprint arXiv:1906.04574
3. Garcia-Teodoro P, Diaz-Verdejo J, Maciá-Fernández G, Vázquez E (2009) Anomaly-based network intrusion detection: techniques, systems and challenges. comput Secur 28(1–2):18–28
4. Bagaa M, Taleb T, Bernabe JB, Skarmeta A (2020) A machine learning security framework for IOT systems. IEEE Access
5. Yamauchi M, Ohsita Y, Murata M, Ueda K, Kato Y (2020) Anomaly detection in smart home operation from user behaviors and home conditions. IEEE Trans Consum Electron 66(2):183–192
6. Li W, Meng W, Au MH (2020) Enhancing collaborative intrusion detection via disagreement-based semi-supervised learning in IoT environments. J Netw Comput Appl 102631
7. Huang HY, Broughton M, Mohseni M, Babbush R, Boixo S, Neven H, McClean JR (2020) Power of data in quantum machine learning. *arXiv preprint* arXiv:2011.01938
8. Alkasassbeh M, Almseidin M (2018) Machine learning methods for network intrusion detection. *arXiv preprint* arXiv:1809.02610
9. Thaseen S, Kumar CA (2013) An analysis of supervised tree-based classifiers for intrusion detection system. In: 2013 international conference on pattern recognition, informatics and mobile engineering, pp 294–299. IEEE
10. Biswas SK (2014) Intrusion detection using machine learning: a comparison study. Int J Pure Appl Math 118(19):101–114
11. Buczak AL, Guven E (2015) A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Commun Surv Tutorials 18(2):1153–1176
12. Abrar I, Ayub Z, Masoodi F, Bamhdi AM (2020) A machine learning approach for intrusion detection system on NSL-KDD dataset. In: 2020 international conference on smart electronics and communication (ICOSEC), pp 919–924. IEEE
13. Kapil D, Mehra N, Gupta A, Maurya S, Sharma A (2021) Network security: threat model, attacks, and IDS using machine learning. In: 2021 international conference on artificial intelligence and smart systems (ICAIS), pp 203–208. IEEE
14. Elmrabit N, Zhou F, Li F, Zhou H (2020) Evaluation of machine learning algorithms for anomaly detection. In: 2020 international conference on cyber security and protection of digital services (cyber security), pp 1–8. IEEE
15. Alraddadi FS, Lago-Fernández LF, Rodríguez FB (2021) Impact of minority class variability on anomaly detection by means of random forests and support vector machines. In: International work-conference on artificial neural networks, pp 416–428. Springer, Cham
16. Ma C, Du X, Cao L (2019) Analysis of multi-types of flow features based on hybrid neural network for improving network anomaly detection. IEEE Access 7:148363–148380
17. Primartha R, Tama BA (2017) Anomaly detection using random forest: a performance revisited. In 2017 International conference on data and software engineering (ICoDSE), pp. 1–6. IEEE
18. Panwar SS, Raiwani YP, Panwar LS (2017) Evaluation of network intrusion detection with features selection and machine learning algorithms on CICIDS-2017 dataset. In: International conference on advances in engineering science management and technology (ICAESMT)-2019, Uttaranchal University, Dehradun, India
19. Stiawan D, Idris MYB, Bamhdi AM, Budiarto R (2017) CICIDS-2017 dataset feature analysis with information gain for anomaly detection. IEEE Access 8:132911–132921
20. Upgrad https://www.upgrad.com/blog/data-preprocessing-in-machine-learning/
21. Analytics Vidhya https://www.analyticsvidhya.com/blog/2020/03/one-hot-encoding-vs-label-encoding-using-scikit-learn/
22. Geek for Geeks https://www.geeksforgeeks.org/data-normalization-in-data-mining/

23. Han J, Kamber M, Pei J (2012) Data Preprocessing, in the Morgan Kaufmann series in data management systems, Morgan Kaufmann, Boston, Data Mining, 3rd edn. ISBN 9780123814791. Elsevier, pp 83–124
24. Thaseen IS, Kumar CA (2017) Intrusion detection model using fusion of chi-square feature selection and multi class SVM. J King Saud Univ-Comput Inf Sci 29(4):462–472
25. Belgiu M, Drăguţ L (2016) Random forest in remote sensing: a review of applications and future directions. ISPRS J Photogramm Remote Sens 114:24–31
26. Han R, Liu P, Wang G, Zhang H, Wu X (2020) Advantage of combining OBIA and classifier ensemble method for very high-resolution satellite imagery classification. J Sens Article ID 8855509, pp 15. https://doi.org/10.1155/2020/8855509
27. Sharafaldin I, Lashkari AH, Ghorbani AA (2018) Toward generating a new intrusion detection dataset and intrusion traffic characterization. ICISSp 1:108–116
28. Kostas K (2018) Anomaly detection in networks using machine learning. Res Proposal 23
29. Syed AH, Khan T, Hassan A, Alromema NA, Binsawad M, Alsayed AO (2020) An ensemble-learning based application to predict the earlier stages of Alzheimer's disease (AD). IEEE Access 8:222126–222143
30. Thaseen S, Kumar CA (2016) Intrusion detection model using PCA and ensemble of classiers. Adv Syst Sci Appl 16(2):15–38
31. Littlestone N, Warmuth MK (1994) The weighted majority algorithm. Inf Comput 108(2):212–261

AutoNav: A Lane and Object Detection Model for Self-Driving Cars

S. Sree Madhumitha, R. Sailesh, A. Sirish, and Jyoti R. Munavalli

Abstract The area of autonomous vehicles is of huge research interest and much has been accomplished in this area. This study involves three aspects: lane detection, object detection, and autonomous driving. Lane detection and object detection has been simulated in the CARLA simulator using TensorFlow and OpenCV libraries of Python. Canny edge detection algorithm and Hough line transform are then used to detect the lane lines. For object detection, image data is collected, labeled manually, and split into test and train data. SSD MOBNET 640 × 640 is used for training the model, and about 75% precision is obtained. Autonomous driving has been implemented in the Udacity simulator using behavioral cloning, a five-layer convolutional neural network (CNN) was used as the model and the data was trained for five epochs with 20,000 steps per epoch. Live predictions are made by the trained model which are used to run the car in autonomous mode.

Keywords Autonomous driving · Lane detection · Object detection · Convolutional neural networks · Image processing · Behavioral cloning · Machine learning

1 Introduction

Autonomous driving is among the most groundbreaking advances in the area of artificial intelligence (AI). It is making an appearance as advanced driver assistance systems (ADAS). An autonomous vehicle is one that can perform all the fundamental functions on its own, without the need for human intervention, through its ability to perceive its environment. The Society of Automotive Engineers (SAE) categorizes automated driving into six levels, ranging from Level 0 (no automation) to Level 5

S. S. Madhumitha (✉) · R. Sailesh · A. Sirish · J. R. Munavalli
BNM Institute of Technology, Bangalore, India
e-mail: ssreemadhumitha@gmail.com

J. R. Munavalli
e-mail: jyotirmunavalli@bnmit.in

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_17

(full automation) [1]. The available sensors are expensive because of which majority of the commercial vehicles offer only Level 1 to Level 2 automation, which requires driver engagement. The autonomous cars of these levels typically include features like adaptive cruise control, lane keeping system, acceleration and braking control, and emergency braking system. These driver-assist functions greatly minimize risks by continuously alerting the drivers and by taking control of the vehicle when absolutely necessary. Autonomous vehicles (AV) have various advantages. It facilitates creation of new jobs in the fields of mobility and logistics, enables convenient and hassle-free transportation, minimizes drivers' stress and fatigue, and boosts their productivity [2]. According to government data, 94% of accidents are due to human behavior [3] and these could be reduced with usage of AVs. This would in turn make roads much safer for pedestrians. AVs can also assist in managing the parking problems, by freeing up to 48% of parking space for users. It can park effectively in a constricted parking space without collision [4]. Object detection is a very important aspect of autonomous vehicles. It makes the car more aware of its environment. Humans are capable of detecting objects and classifying them and yet make errors. Allowing a car to identify the objects around it and classify them can help the car to make decisions to avoid potential crashes and thus increasing the overall safety on the road. There are five main components of autonomous vehicle navigation perception, localization and mapping, path planning, decision making, and vehicle control [5]. Perception utilizes various sensors like ultrasonic sensors, LIDAR, RADAR, and cameras to scan and monitor the environment and obstacles around the vehicle much like human vision and other senses. Problems such as obstacle detection (vehicles, pedestrians, traffic signals, and so on), lane detection, and scene recognition can be addressed by using deep learning or traditional computer vision techniques [6]. Localization and mapping algorithms define the position of the vehicle in the world and maps the environment based on sensor data and perception output. There are various techniques for a vehicle to position itself such as active sensor-based localization like simultaneous localization and mapping (SLAM), passive sensor-based localization like extended Kalman filter, global positioning system (GPS), inertial sensors, odometry and landmark detection, and particle filters (landmark detection and Dara association), Path planning is used for charting a trajectory between two points using perception, localization, and mapping. Decision making is in charge of computing the best possible path based on the available pathways, the present vehicle status, and information about its surrounding environment. The vehicle control follows the trajectory by calculating the suitable vehicle commands such as torque, steering angle, and acceleration value.

This paper largely deals with the perception section of autonomous vehicles. The organization of the paper is as follows: Sect. 2 presents the methodologies used for this study. Section 3 presents results and Sect. 4 presents the discussion. Conclusion is presented in Sect. 5.

2 Materials and Methods

The simplest approach to the defined problem is the usage of image processing and deep learning techniques [1]. To achieve the objective, this study was divided into three parts: lane detection, object detection, and autonomous driving.

2.1 Lane Detection

Figure 1 shows the lane detection methodology. The image data is read from RGBA image data provided by the RGBA sensor of the CARLA simulator. The RGBA image is converted to grayscale. All images have noise, and it is in the form of

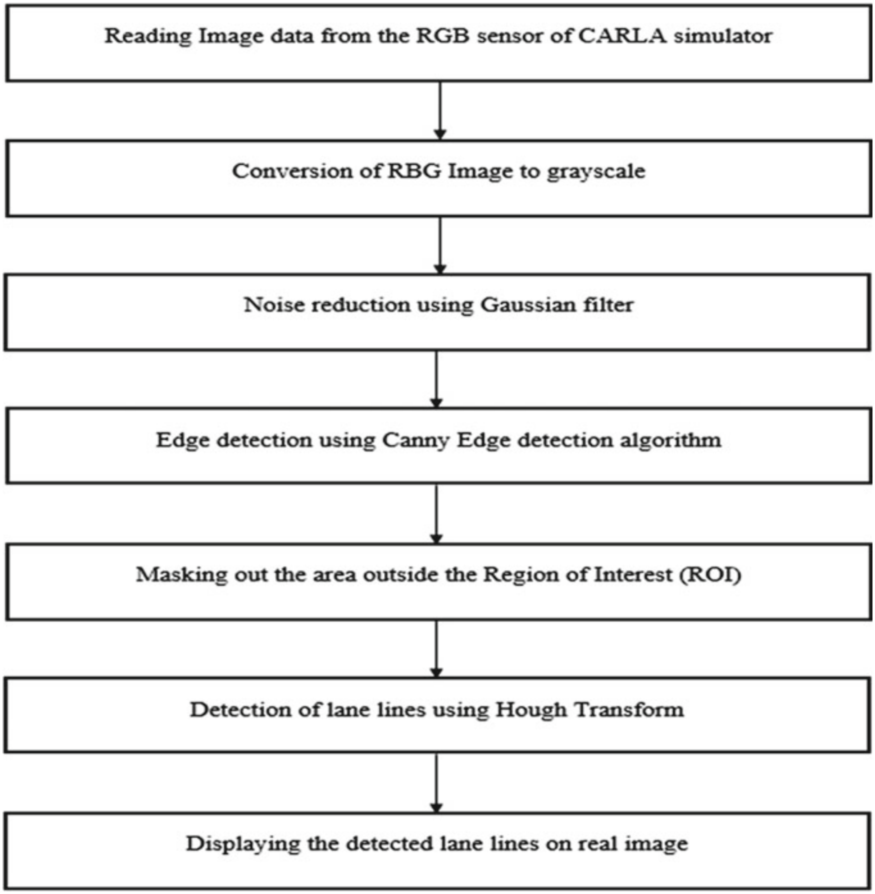


Fig. 1 Lane detection methodology

Gaussian noise or salt and pepper noise. Image noise can create false edges and effect edge detection [3]. We use the Gaussian blur function of the Python OpenCV library which is a very effective image smoothing algorithm [4] to reduce the Gaussian noise. This noise reduced image which is fed to an edge detection algorithm. The Canny edge detection algorithm is used for edge detection. Edge detection is performed to identify the strongest gradients in the image. The grayscale image is taken as the input and an output is generated, which displays the outline of white pixels. The white pixels traced correspond to the discontinuity in the brightness of the pixels. This algorithm operates in a multi-staged approach. The image is first smoothed out by using Gaussian convolution. The smoothed image is then subjected to a simple 2D first derivative to emphasize the parts of the image with high first spatial derivative. In the gradient magnitude image, the edges are displayed as ridges. This algorithm tracks the top of the ridges and converts all the pixels that are not on the top of the ridge to zero, resulting in a narrow line in the output that represents the edge [5]. Next, the region of interest (ROI) is chosen. A mask of the same dimension of the selected road image is created. A bitwise AND operation is performed between each pixel of the Canny image and the mask. This is done to mask out the unwanted area of the image. The ROI is calculated manually through trial and error.

The Hough line transform is a feature extraction technique used to detect straight lines. It converts the pixels in the edge detected image to polar form. Lines in the Hough space correspond to points in the image. The intersection of lines in the Hough space is equivalent to a line in Cartesian space. The lane lines are located using this technique from the pixel output of Canny edge detector. The final step is to blend the two images, the original image from the camera feed of the car and the image that contains the detected lane lines [6].

2.2 Object Detection

Object detection is performed in three stages: acquisition of data, training the model, and testing of the model. There are many popular widely used image detection techniques like YOLO, Faster R-CNN, SSD mobile nets, etc. YOLO struggles to detect close objects and comparatively, has more recall and more localization error. Methods like Faster R-CNN require multiple passes of the same image to extract all the objects which makes R-CNN a bad choice especially in the cases where we need to detect multiple objects in the same frame from multiple frames a second [7]. In this study, SSD mobile nets are implemented. SSD mobile nets offer a good balance between speed and utilization of computer power. Figure 2 shows the flow diagram for object detection.

The first step is image data collection and for this Carla simulator is used. A car is simulated and spawned with a RGB sensor attached to it. The car is allowed to move around in the simulated environment. The simulated environment consists of many vehicles and pedestrians and is referred as Non-Participant Characters (NPC). During the movement of the car, it collects images at about one frame per second.

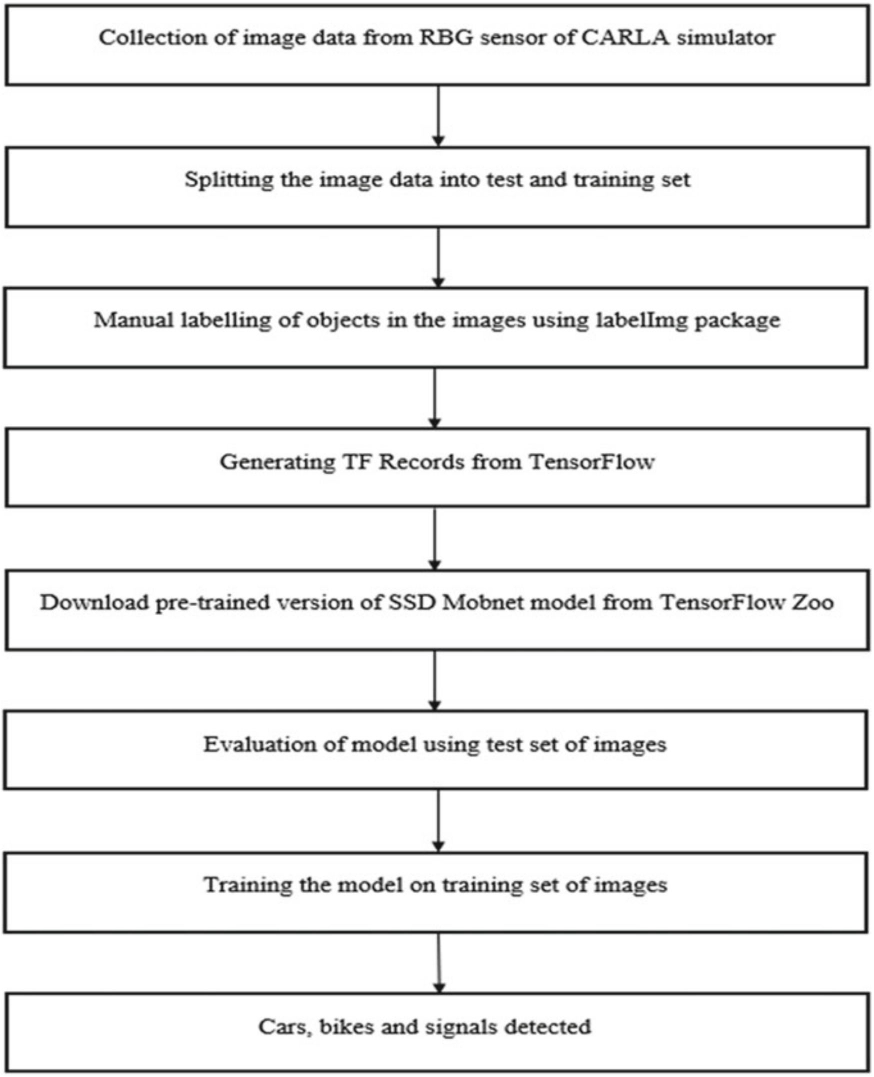


Fig. 2 Object detection methodology

Collecting images at a higher rate would be unnecessary as this would result in a vast amount of similar looking images. A sample size of 50 images is selected based on labels—car, bike, signal, pole, and pedestrian. These images are then manually labeled by drawing bounding boxes around them using the Python labelImg package. The bounding boxes are stored in XML files in the form of coordinate values. Each bounding box can be described as a collection of four coordinates which determine where the box is located in the image. The XML files are saved with the same name as the image itself, so while generating the tfrecords, TensorFlow knows which box to associate with which image. After drawing bounding boxes around a sufficient number of images, the images are separated into a test set and a train set in a 4:1 ratio. Since SSD mobile net is used for this part of the study, object detection API provided by TensorFlow is used. This API requires labels to be provided in the form of protocol buffers. Protocol buffers are a language-neutral, performance-neutral, and flexible mechanism used for serializing structured data developed by Google [8]. It also requires image data to be supplied to it in the form of TensorFlow records (or tfrecords). A Python script is developed to convert labeled data into test and train records. A pre-trained version of the SSD MOBNET (SSD MOBILE NET V2 FPNLITE) is available for download from the TensorFlow model zoo. Initially, the model was developed using an SSD MOBNET 320 $\times$ 320 model and later using SSD MOBNET 640 $\times$ 640 model. This model compresses input images to a square of 320 pixels and 640 pixels, respectively, on each side. Both the models were trained on a set of 50 images for 2000 steps. This trained model detects cars (which includes cars, trucks, and any other four wheelers), bikes (which includes both bikes and bicycles), pedestrians, signals, and poles. The 640 $\times$ 640 model is trained again with a larger dataset. This time, the image set was collected from a variety of weather conditions. This model showed decreased precision and recall values and increased loss. However, this model showed higher confidence levels in predicting cars and bikes in practice.

2.3 Autonomous Driving

The usage of convolutional neural networks (CNNs) is widespread in the world of autonomous driving as it allows mapping of image data that is provided by a front-facing camera to steering commands in a self-driving car. This approach means that a system can learn to navigate itself with minimal human training data, on roads with or without lane markings [9]. We use the concept of behavioral cloning for this. The groundwork for this topic was laid over 15 years ago by the Defense Advanced Research Project Agency [DARPA] and the project was called as DAVE—DARPA Autonomous Vehicle [10]. Several improvisations have been made to this framework in recent years. This method has been called the DAVE-2 system [11].

Figure 3 shows the methodology for autonomous driving. The Udacity simulator comes with three front-facing cameras attached to the vehicle. One camera is at the hood of the car, one is to the front-left wheel of the car and the last camera is at the front-right wheel of the car. These three cameras are positioned as proposed by Nvidia's Dave-2 model. These three cameras collect data at the rate of about ten frames per second. These collected images are stored in the computer memory. Along with these images, information about the throttle, brakes, reverse (reverse velocity), and steering angle are noted down at the instance at which the image is taken. These values are noted down in a csv file. The car was driven for around 20 laps around the track during which it collected about 41,000 images which created 13,695 unique data-points.

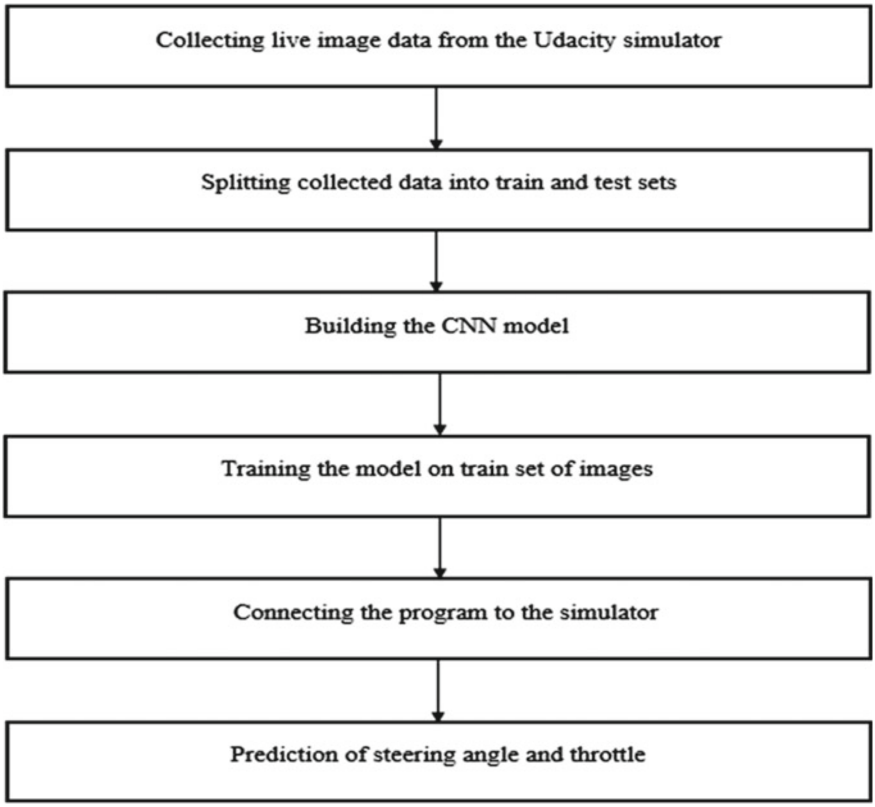


Fig. 3 Autonomous driving methodology

It is observed that the track has a lot of left turns and straight stretches of road and very few left turns. The simplicity of the track allows full throttle for the whole lap. This data is divided into train set and a test set with a ratio of 4:1. The convolutional neural net we used for this application consists of five two-dimensional convolution layers, one image normalization layer, and four fully connected layers with 100, 50, 10, and 1 neurons in each layer, respectively. The input layer is the image data flattened out. In addition to these layers, there is a dropout layer which removes 50% of the data. Since most of the steering angle values are negative in our dataset, the dropout layer helps prevent overfitting the data. The model was built using Python's TensorFlow library and their Keras API. The optimizer used in this model is the Adam optimizer and the activation function used is exponential linear unit (ELU).

A socket connection is established between the program, which is the server and the simulator which is the client. This connection allows two-way transmission of data. The car sends image data from the three cameras at around a rate of ten frames per second, per camera. Higher FPS is not a necessity as it would result in a lot of similar looking images with not much valuable information. These images are then fed into the trained model which predicts the required steering angle. The max speed of the car is limited to 25 kmph on this particular track.

3 Results

The proposed study was implemented using the Python 3.7 with libraries OpenCV, TensorFlow, TensorFlow object detection API, pre-trained object detection model from TensorFlow model zoo and Microsoft Visual C++ tools. This study required to simulate the real-time environment to test the algorithms and methodologies. So, CARLA simulator and Udacity simulator were used. Both are open-source simulators.

The Carla simulation platform allows customizable specification of sensor suites, environmental conditions, control of static and dynamic actors, generation of maps, etc. It runs on localhost port 2000 by default. The simulator also comes with a Python library which can be imported in a Python script which allows to connect to the simulator. On successfully connecting to the simulator, map can be selected from available maps. After loading the required map, the weather conditions on the map can be changed. There is also an option to spawn up to 200 cars and pedestrians at random points in the map. The user can have complete control over a single car in the simulation. This involves being able to send throttle and steering control to the car through code. The user is also allowed to attach multiple cameras at various locations on the car. Udacity is a self-driving car simulator built with Unity. This simulator consists of a single track which replicates a one-way road with lane markings on either side. There is no vehicle and pedestrian traffic. We present the results according to the methodologies: lane detection, object detection, and autonomous driving.

3.1 Lane Detection

The video or the image feed comes from the simulator through an RGB camera attached to the hood of a car. Each frame is then fed as input to the algorithm that detect the lanes. The output of lane detection is as shown in Fig. 4. The lane lines are successfully detected in real time.

3.2 Object Detection

The images being fed were of size 1920 × 1080 (full HD 1080p). We first implemented object detection using the 320 × 320 SSD MOBNET. It was unable to detect pedestrians accurately as this model compressed the data to a square of 320 px per

(a)Canny output

(b) Cropped Image

(c)Hough Transform Output

(d)Blended Image

Fig. 4 Outputs **a-d** of the lane detection method

side and a lot of data was lost. This model was trained on 50 images with 2000 steps, it has a training loss of 0.29 and a validation loss of 0.80 by the end of training. The model had an average precision of 0.70 at IoU = 0.50 and an average recall of 0.29. It was concluded that 320 × 320 model does not predict the objects well.

We then changed the model to SSD MOBNET 640 × 640. This model compresses images down to a square of 640 px per side and not much data was lost. This model was trained again on about 50 images for 2000 steps. The training loss and validation loss almost decreased by half, to 0.096 and 0.4209, respectively. The precision values increased to 0.75 at IoU = 0.5, and the recall value increased to 0.419. These results are much better compared to 320 × 320 model. The above-mentioned training process was done on a small set of images to test the performance of the models. And the weather conditions that were simulated in the simulator was daytime conditions.

Figures 5a and b shows the predictions made by both the models and the manually drawn bounding boxes. Figure 6 shows the loss over time for 320 × 320 model compared to 640 × 640 model, trained for 2000 steps.

Next, we trained the 640 × 640 SSD MOBNET model on a much larger dataset of images of about 500 images and the weather conditions in the simulator were changed to an overcast sky. It is found that with this large of a dataset, the training loss value is 0.14 and the validation loss is 0.98. It can be observed that the training value starts to creep up after around 1600 steps of training. The precision for this particular model which was trained with a larger dataset dropped off to around 60% and the average recall dropped off to about 41% at IoU = 0.50. With a larger training dataset, we expected higher accuracy values, however, the model ended up

Fig. 5 **a** Image on the left shows the predictions made by the model and the image on the right shows the bounding boxes that were drawn manually for 320 × 320. **b** Image on the left shows predictions made by the model and the image on the right shows the bounding boxes that were drawn manually for 640 × 640

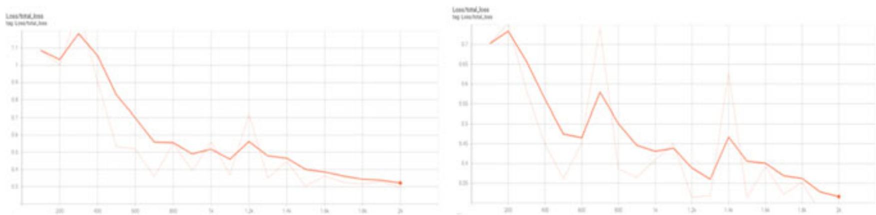


Fig. 6 Total loss on training 320 × 320(left) and 640 × 640(right) SSD MOBNET model for 2000 training steps (50 images)

performing worse than the previous time. This is possibly because of two issues—the labeling of certain images such as poles was done inconsistently which might have reduced the accuracy of detection for poles which resulted in a drop-off of the total accuracy. Another reason as to why there was a drop-off is that the images that were collected on this run were collected with bad weather conditions simulated in the simulator. With an overcast sky, the absence of sun caused a drop-off in the light which affected the performance of the model. It can be seen in Fig. 7 that model shows higher confidence values of signals, cars and bikes, but performs extremely poorly on pedestrians and poles.

Fig. 7 The image on the left shows the predictions made by the model and the figure on the right shows the bounding boxes that were drawn manually

3.3 Autonomous Driving

Higher accuracy values are expected from algorithms such as Faster R-CNN which would mean a sacrifice to the number of frames per second that can be handled. With this model, we were able to achieve about ten frames of live classification per second with the simulator running which used up more of the computer's resources and about 25 frames per second of classification done without the simulator running. This leaves much scope for improvement in the future, both in terms of better labeling the images or collecting a wider set of images and in terms of picking a better model for the task. Figure 8 shows the summary of the driving model. Figure 9 depicts the simulator where steering angle and throttle are displayed. Figures 10 and 11 show the epoch accuracy and losses.

The model is then trained on the dataset for five epochs with 20,000 steps per epoch. Epoch accuracy and loss for 14,000 data-points and 34,000 data-points are compared.

4 Discussion

We build a CNN that can process image data and map it to steering controls to make sure that a car avoids possible obstacles and learns to drive safely around a track. We use the concept of behavioral cloning, which is a much-advanced form of pattern matching.

One of the main reasons for choosing behavioral cloning is that it doesn't necessarily need a large amount of data to be able to make accurate predictions. It is also possible to attach this to a reward system when the car collides or crosses its designated lane. Another widely used technique in the area of autonomous driving is reinforcement learning. The difficulty with implementing a reinforcement learning model is to come up with a concrete reward function. The number of variables and the size of the learning environment itself in the case of self-driving cars makes it difficult to implement reinforcement learning. In addition to that, it is also an extremely

Model: "sequential"

Layer (type)	Output Shape	Param #
lambda (Lambda)	(None, 66, 200, 3)	0
conv2d (Conv2D)	(None, 31, 98, 24)	1824
conv2d_1 (Conv2D)	(None, 14, 47, 36)	21636
conv2d_2 (Conv2D)	(None, 5, 22, 48)	43248
conv2d_3 (Conv2D)	(None, 3, 20, 64)	27712
conv2d_4 (Conv2D)	(None, 1, 18, 64)	36928
dropout (Dropout)	(None, 1, 18, 64)	0
flatten (Flatten)	(None, 1152)	0
dense (Dense)	(None, 100)	115300
dense_1 (Dense)	(None, 50)	5050
dense_2 (Dense)	(None, 10)	510
dense_3 (Dense)	(None, 1)	11

Total params: 252,219
Trainable params: 252,219
Non-trainable params: 0

Fig. 8 Summary of autonomous driving model

data hungry algorithm. Behavioral cloning is a tried and tested and well researched method for implementing driving assistance features and has been known to translate well to the real world. We have used the Udacity simulator to achieve this.

There are many approaches to lane detection and the popular ones among them are by using convolutional neural nets [9] and an approach using image processing techniques [10]. We have made use of the image processing approach to implement this model. While this approach may be less reliable compared to the CNN-based approach, it is computationally light weight and does a great job on roads where the lane markings are clear. Another downside to using image processing techniques is that the algorithm performs poorly in low light conditions and in places where there are lots of shadows. Both these approaches have similar steps for lane prediction and

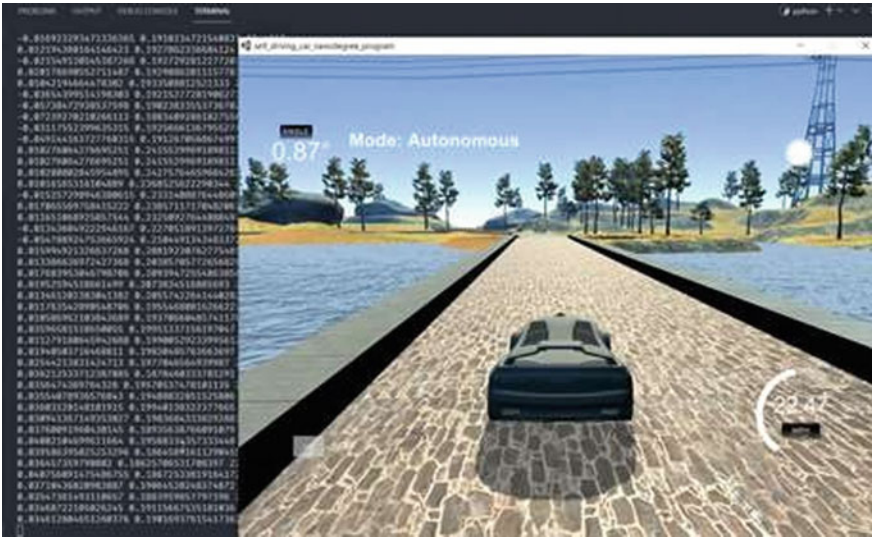


Fig. 9 Simulator being run in the foreground with the program sends values of steering angle and throttle in the background

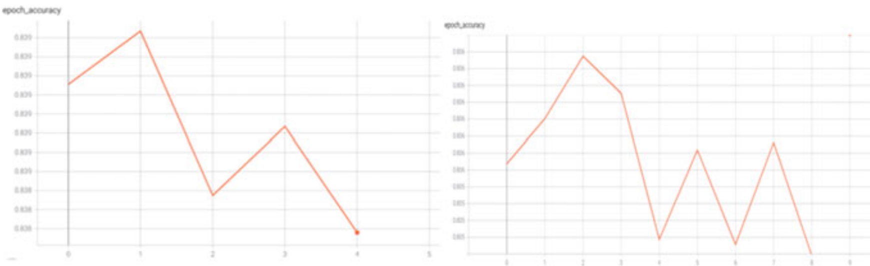


Fig. 10 Epoch accuracy for 14,000 data-points (left) and 34,000 data-points (right)

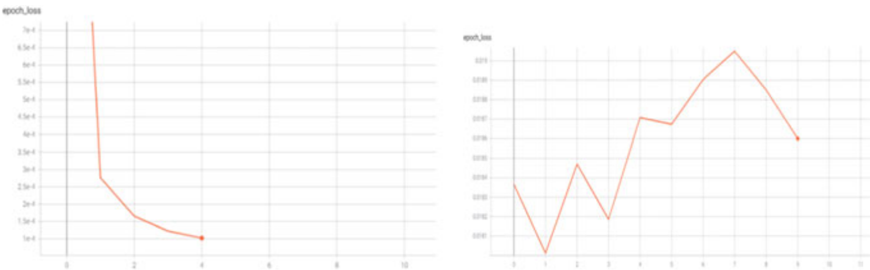


Fig. 11 Epoch loss for 14,000 data-points (left) and 34,000 data-points (right)

lane fitting. Where the CNN approach excels is in the steps of semantic segmentation and clustering. Another problem while implementing the image processing algorithm is in the selection of a region of interest. This is solved by using a dynamic region of interest selection algorithm. Here, the region of interest is selected by using the curvature of each point on the edge of the road and the maximum safe distance which is based on the speed of the vehicle in the previous frame. This algorithm is found to have an accuracy of close to 99% [11]. As a future scope, this model can further be improved by adding a lane departure warning system [12, 13].

Lane detection during daytime in the simulator showed poor results due to shadows and broken lines. It was still able to detect lanes but there were problems when it had to detect broken white lines. Lane detection also showed poor results in rainy weather conditions due to reflections from puddles of water. Compared to that, lane detection at night time was much more consistent.

Object detection mainly consists of three steps: collecting data, choosing a model, and training the data. There are different models which can be used to train the data such as Fast R-CNN, Faster R-CNN, Region-based convolutional neural networks (R-CNN), Region-based fully convolutional networks (R-FCN), Histogram of oriented gradients (HOG), You only look once (YOLO), Spatial pyramid pooling (SSP-net), and Single shot detector (SSD). SSP-net is comparatively very slow hence couldn't be used for live detection [14]. YOLO though it is faster the accuracy is below subpar and also struggles with detecting small objects as it predicts only one class in a given grid [15]. Fast R-CNN has a good accuracy rate but, however, requires more time [14, 16].

SSD MOBNET has a good speed versus accuracy trade-off and was also easy to train. It uses small convolution filters to predict object classes and bounding box locations for different aspect ratios [17]. Hence, we implemented this model to detect objects. There are four different models under this—SSD MobileNet v2 320 × 320, SSD MobileNet V1 FPN 640 × 640, SSD MobileNet V2 FPNLite 320 × 320, and SSD MobileNet V2 FPNLite 640 × 640.

5 Conclusion

Autonomous vehicles have been a major area of research interest for the past few decades. Several important factors contribute in the creation an efficient and safe real-world autonomous vehicle. The perception of one's surroundings remains the most difficult barrier to a dependable, smooth, and safe driving. Our project is an attempt to understand a few important technologies that will aid in the development of a self-driving car. Our paper deals with the major aspects of the perception system of AV's—lane and object detection.

We wish to extend the scope of this project by integrating the above features and adding additional features. We aim to develop a cohesive working model of an advanced driver assistance system which includes which includes features like collision warning and lane departure alert system that can be implemented in the real world.

References

1. SAE (2021) On-road automated vehicles standards committee. Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. J3016_202104
2. Litman T (2015) Autonomous vehicle implementation predictions: implications for transport planning
3. Singh S (2015) Critical reasons for crashes investigated in the national motor vehicle crash causation survey
4. Hancock PA, Nourbakhsh I, Stewart J (2019) On the future of transportation in an era of automated and autonomous vehicles. Proc Natl Acad Sci 116(16):7684–7691
5. Van Brummelen J, O'Brien M, Gruyer D, Najjaran H (2018) Autonomous vehicle perception: the technology of today and tomorrow. Transp Res Part C: Emerging Technol 89:384–406
6. Ni J, Chen Y, Chen Y, Zhu J, Ali D, Cao W (2020) A Survey on theories and applications for self-driving cars based on deep learning methods. Appl Sci 10(8):2749
7. Samak TV, Samak CV, Kandhasamy S (2020) Robust behavioral cloning for autonomous vehicles using end-to-end imitation learning. 04767
8. Hall G, Das A, Quarles J, Rad P (2020) Studying adversarial attacks on behavioral cloning dynamics. In: 2020 IEEE 32nd international conference on tools with artificial intelligence (ICTAI), Baltimore, MD, pp 452–459
9. Wang W, Lin H, Wang J (2020) CNN based lane detection with instance segmentation in edge-cloud computing. J Cloud Comp 9:27
10. Somasundaram G, Kavitha, Ramachandran KI (2011) Lane change detection and tracking for a safe-lane approach in real time vision based navigation systems
11. Hu J, Xiong S, Sun Y, Zha J, Fu C (2020) Research on lane detection based on global search of dynamic region of interest (DROI). Appl Sci 10(7):2543
12. Hsiao P, Yeh C, Huang S, Fu L (2009) A Portable vision-based real-time lane departure warning system: day and night. IEEE Trans Veh Technol 58(4):2089–2094
13. Jung CR, Kelber CR (2005) Lane following and lane departure using a linear-parabolic model. Image Vision Comput 23(13):1192–1202
14. Girshick R (2015) Fast R-CNN arXiv:1504.08083[cs.CV]
15. Gao W, Tang J, Wang T (2021) An object detection research method based on CARLA simulation. In: 2021 journal of physics: conference series, the 2nd international conference on internet of things, artificial intelligence and mechanical automation (IoTAIMA 2021) vol 1948, pp 14–16, Hangzhou, China
16. Zhao Z, Zheng P, Xu S, Wu X (2019) Object detection with deep learning: a review. IEEE Trans Neural Netw Learn Syst 30(11):3212–3232
17. Sheri R, Jadhav N, Ravi R, Shikhare A, Sannakki S (2018) Object detection and classification for self-driving cars

Using Natural Language Processing to Understand Reasons and Motivators Behind Customer Calls in Financial Domain

Ankit Patil, Ankush Chopra, Sohom Ghosh, and Vamshi Vadla

Abstract In this era of abundant digital information, customer satisfaction has become one of the prominent factors in the success of any business. Customers want a one-click solution for almost everything. They tend to get unsatisfied if they have to call about something which they could have done online. Moreover, incoming calls are a high-cost component for any business. Thus, it is essential to develop a framework capable of mining the reasons and motivators behind customer calls. This paper proposes two models. Firstly, an attention-based stacked bidirectional long short-term memory network is followed by hierarchical clustering for extracting these reasons from transcripts of inbound calls. And, secondly a set of ensemble models to classify the motivators of the inbound calls. It is capable of detecting factors that led to these calls. Extensive evaluation proves the effectiveness of these models.

Keywords Financial text summarization · Financial text clustering · Financial text classification · Natural language processing

1 Introduction

Phone calls are one of the main channels through which customers interact with organizations. Customers call either seeking answers to their queries or for getting a service request fulfilled. For contacting customer care, typically customers need to select the right options in the interactive voice response (IVR) after going through the menu, and there is generally a caller queue resulting in a hold time. Hence, phone calls tend to be a slower medium of communication by design compared to mediums like chat, virtual assistants (VA) and search.

In most cases, phone interactions result in suboptimal experience from both organization and customers' points of view, due to the time, it takes to get the information. It also results in higher operational cost of call centres.

A. Patil (✉) · A. Chopra · S. Ghosh · V. Vadla
Fidelity Investments, AI CoE, Bengaluru, India
e-mail: ankitpatil4@gmail.com

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_18

There has been a push for digitization of services to enable and empower customers towards self-service. This will enable faster access to information, efficient service request resolution for the customers while reducing the cost of service for the organizations.

For effective digitization of services, it becomes essential to understand the reasons behind the customer calls and the gaps in the current digital experience. In this paper, we propose a framework to understand the reasons and motivators of these calls using machine learning and natural language processing (NLP) techniques. These calls are inbounded to fidelity investments[1] and are specific to the financial domain.

Understanding Call Reasons

We summarize the call conversation into customer intents of length up to 6 words. The abstractive summarization on the stored transcripts of the calls, provides the reason behind the call. We have used the long short-term memory network [12] with attention [29] to generate the summary intents.

Understanding Call Motivators

In this work, we have researched solutions about "why the customer had to call in": is it because they are not tech-savvy, or they were not able to find the information they are looking for or something for which there is no digital solution available at present. Innovations in this space are a win–win for both the customer (information at fingertip) and for business (cost saving by reducing inbound calls).

Challenges

Data scarcity is the biggest challenge when coming up with a solution for the stated problem. A dataset where each of the calls has been tagged with the primary reason and motivation of the call would be the ideal dataset for coming up with solutions. Multiple transfers during the call make it difficult to always decipher the actual reason for the call. Also, there can be many latent motivations for the call, as this is very subjective at times, e.g. "*I need help with the website*" or "*I have attempted to do this online but was unable*". The former says the customer was not aware of what to do on the website, but the latter informs us that the customer had tried it and was unable to do it online. This may be because of some misleading information or some issue with the pertaining account.

We made use of customer representative summaries that were available for a fraction of calls to generate the data needed for training the model capable of detecting reasons behind calls. For call motivators, we took the help of the domain experts and representatives who listened to these calls and assigned one label (primary motivator) to every call. Coming up with the labels was an iterative process. After few iterations, we decided to focus on 11 call motivators which were of major interest to the stakeholders. As this list is not exhaustive, we introduced an "other" label as one broad category which can be mined later for any more categories.

[1] https://www.fidelity.com/.

Table 1 Call motivators and their corresponding occurrence

Motivators	Occurrence (%)
M1	22.94
M2	1.21
M3	2.47
M4	1.02
M5	3.91
M6	18.44
M7	6.68
M8	13.90
M9	1.67
M10	14.09
M11	7.42
Other	6.24

Contributions. Our contributions are as follows: i) We developed a summarization model and a subsequent clustering model to understand reasons relating to these calls ii) We developed a classification model to comprehend the factors which led to these calls.

Combining findings from both models gave us a clear view of the avenues, where digitization can be done or improved. We will be describing the dataset in the next Sect. 3 and followed by the solution methodology Sect. 4. We will shed some light on the experimentation and results Sect. 5 after the methodology section. We will discuss our findings Sect. 6 and conclude Sect. 7 the paper by narrating the future enhancements to this framework.

Problem Statement

Our problem statement is two folds. Given a set $C = \{ c_1, c_2, c_3 \dots c_n \}$ of calls specific to financial domain, we need to (i) identify call reasons and cluster them. Call reasons are an open set of customer issues (like "adding a bank account", "password reset", "transfer of assets" and so on) (ii) classify these calls into 12 pre-defined categories of motivators mentioned in Table 1.

2 Related Works

Some of the conventional extractive text summarization techniques include Text-Rank [19] (a graph-based ranking model) and LexRank [8]. Text-Rank is an unsupervised method for extracting keywords and sentences. LexRank is a graph centrality-based model to score sentences.

Some of the newer text summarization techniques include genetic algorithm-based approach MUSE [15] and linear programming-based approach POLY [16]. Shi et al. in their survey paper [25] described various neural network-based abstractive summarization processes. They further developed and open-sourced an abstractive summarization library named neural abstractive text summarizer (NATS) toolkit. Gupta et al. in their paper [10] surveyed several extractive summarization techniques. In the paper [30], Verma et al. described an ontology-based summarization model tuned specifically for medical documents. They used WordNet and Unified Medical Language System to revise keywords for ranking sentences. The Financial Narrative Summarization Shared Task (FNS 2020) [7] was organized by El-Hai et al. This task was about summarizing annual reports of organizations listed in London Stock Exchange, UK. They discussed the efforts of 9 teams. The best performing team SUMSUM comprising Zheng et al. [33] parsed these financial reports into sections based on table of contents (ToC) and applied BERT [6] to each of them. They achieved a F1 score of 0.306. Singh [26] ranked second by achieving a F1 score of 0.289 using pointer network [24] and T5—Text-to-Text Transfer Transformer [21]. Azzi et al. of team FORTIA [2] ranked third. They used rule-based approaches to extract ToC. They also used a convoluted neural network-based binary classifier to identify candidate summaries. Their F1 score was 0.274. A goal guided summarization [1] model had been described by Agrawal et al. [1] in their paper. They narrated how they summarized annual reports of organizations using hierarchal neural network models to infer whether to buy or sell a stock.

So far, we discussed various research works on financial text summarization. Now, we will discuss prior works related to financial text classification. Ciravegna et al. [5] developed a financial text classification system called FACILE. This system was flexible and it worked for corpora of 4 different languages, namely Italian, English, German and Spanish. Zhao et al. [32] described how they used partial information to classify Chinese Financial News. Yang et al. [31] in their paper discussed how much explainable transformer-based financial text classification models are. Various pre-processing techniques for classifying financial texts had been described by Sun et al. [27] in their paper.

It is quite interesting to note that while a lot of work has been in the area of financial text processing, only a few of them ([11] and [4]) tried to understand the intent behind inbound calls specific to the financial domain. Moreover, none of them performed abstracting summarization followed by hierarchical clustering and classification of customer calls related to the financial domain to analyse the reasons and motivators behind these calls.

3 Data

Data used for training Call Reason Models

We first pre-process and clean call transcripts. We remove the system noise, the transcription-induced noise and masked tokens. We then remove the pleasantries and other non-informative phrases. We also perform case normalization and contraction replacement. Customer care representatives summarize the conversations they had during a call and notes them down at the end of the call. These summary notes are only available for a fraction of calls (˜20%). We call these summaries repnotes. Next, we clean repnotes by performing steps like lower case conversion, removal of masked tokens, replacement of contraction and non-informative phrases. Details of the pre-processing steps can be found in Table 2.

We looked at the length distribution of cleaned call transcripts and repnotes. About 90% of the call transcripts were of 425 words or less, with a long tail for the remaining 10%. Repnotes with a length of 6 or less constitute 35% of the entire repnotes data. About 40% of them are of length between 7 and 17, and the rest have 18 words or more. Since we are trying to generate concise customer intents from the calls, we decided to use the repnotes which were in the first bucket. We divide the data into the train (70%) and validation (30%) set. We considered all the calls with less than or equal to 425 tokens in transcripts, where repnotes of length 6 or less were present for modelling. This was close to 1.9 million interactions.

Data used for training Call Motivator Models

Data Labelling Exercise. Mining motivators from customer calls is a hard task mainly due to the subjectivity associated with it. A labelling exercise was designed to know what kind of different call motivators can be present in the financial domain. It was an iterative process executed by a variety of customer representatives. Finally, 11 categories are selected but those did not cover 100% of the calls. Thus, an "Other" label was introduced. In all, we have 12 categories of business interests (primary call motivators) distributed over 6000 inbound calls. The data set is highly imbalanced, with the most frequent category "*Attempted to do something online but was unable to*" had an incidence of 23% and most infrequent category "*Following up a communication*" had incidence of only 1%. The details mentioned in Table 1. Due to internal security reasons, we mask these motivators as M1, M2 … M11. We used 70% of this data for training and 30% for validation.

Data Discovery. With labelled data available, this becomes a text classification problem for us. The call interaction was available to us in the form of transcripted

Table 2 Pre-processing details for call reason models

Pre-processing (normalization) step	Call transcripts	Repnotes
Call transcript data contain system messages like "*party has left the session*"	Yes	No
Lower case conversion	Yes	Yes
Replacing the masked tokens	Yes	Yes
Removal of non-vocalized noise transcription	Yes	
Contraction replacement	Yes	Yes
Removal of non-informative prefixes like "customer contacted" or "customer asked"	No	Yes

text. These transcripts (noisy at times) were not enough. So, we have explored other data streams around an inbound customer call.

Most of the customers leave trails of data before and after the call:

1. Before calling, customers try to find information online, so they visit the website (clickstream data).
2. When they call, they need to state their need (whisper) to help us route the call.
3. During the call, they might browse through the website (clickstream).
4. After the call, we have the transcript of the call.
5. The representative writes a note/summary of the call for future references.
6. We also have a model, which explains the reason for a call and further generates a summary of the call interaction.

4 Methodology

This work is divided into two major parts, identifying the reasons behind the calls and classifying the call motivators. We initiate by describing the first of the two major parts of the framework.

Understanding Call Reasons

Call transcripts are stored at the utterance level, where every entry is the transcription of what the agent or customer has said. Calls have an introduction, pleasantries and informal conversation messages along with the actual business conversation. This makes transcripts lengthier, where customer intent is hidden somewhere in the body of the conversation. To be able to understand the reason behind the call, we summarize the call into short customer intent. We use the transcript and human-generated summaries to create the call reason model.

Customer calls are transcribed and stored in the form of text. We use these transcripts with repnotes. Reps inadvertently end up introducing a lot of variations and subjectivity into the repnotes even if they mean the same thing. For example, "*customer contacted to get account reset*" and "*customer asked for the help with password reset although worded differently*", mean the same.

We have used transcribed calls and the repnotes as the training data for the intent generation model. Call transcripts are used as input to the model, and repnote is used as summary intent or the reason behind the call. Since we are looking for short customer intents, we limit our data to only inbound calls, where repnotes were written with 6 words or less.

Repnotes are abstractive summarizations of the conversation that happened in customer calls. We use LSTM [12] architecture with attention [29] to model this. Attention particularly helps here to handle the transcripts well which otherwise be difficult due to their length. It is a two-layered stacked bi-LSTM sequence-to-sequence with attention architecture. We have not used any pre-trained embedding due to the nature of the underlying data. Instead, we are learning the embedding while training for this task.

The model hit the early stopping with a validation categorical cross-entropy loss of 1.884 at the 28th epoch. This model is used to generate the customer intent behind all incoming calls and chat transcripts irrespective of whether they have repnotes or not. These generated repnotes look much more like short customer intents and explain the reason behind the calls. The model summary has been mentioned in Fig. 1. Other details include: latent dimension = 300, embedding dimension = 150, size of vocabulary for repnote = 18,048, size of vocabulary for input transcript = 31,363, optimizer = rmsprop and loss = sparse categorical cross-entropy.

These generated intents are sufficient to explain the reasons behind the calls. Every year, our organization fidelity investments receives millions of customer calls. A framework to digest this data easily is immensely useful. We added a layer of clustering to make this data more actionable. We take all the generated customer intents and apply some pre-processing and normalization steps. We replace acronyms with their definition, remove repetitive words and remove any customer-specific information which may have been captured in the generated summary. Further, we convert these cleaned customer intents to their equivalent mathematical representation or phrase embeddings. We used sentence transformers [22] to come up with phrase embeddings while using the RoBERTa [18] embeddings at the back. Finally, we use agglomerative clustering as mentioned in [3] to group these call reasons into homogeneous buckets.

```
Model: "model"
____________________________________________________________________________________________________
Layer (type)                    Output Shape         Param #     Connected to
====================================================================================================
input_1 (InputLayer)            [(None, 425)]        0
____________________________________________________________________________________________________
embedding (Embedding)           (None, 425, 150)     4704450     input_1[0][0]
____________________________________________________________________________________________________
bidirectional (Bidirectional)   [(None, 425, 600), ( 1082400     embedding[0][0]
____________________________________________________________________________________________________
input_2 (InputLayer)            [(None, None)]       0
____________________________________________________________________________________________________
bidirectional_1 (Bidirectional) [(None, 425, 600), ( 2162400     bidirectional[0][0]
____________________________________________________________________________________________________
embedding_1 (Embedding)         (None, None, 150)    2707200     input_2[0][0]
____________________________________________________________________________________________________
tf_op_layer_add (TensorFlowOpLa [(None, 300)]        0           bidirectional_1[0][1]
                                                                 bidirectional_1[0][3]
____________________________________________________________________________________________________
tf_op_layer_add_1 (TensorFlowOp [(None, 300)]        0           bidirectional_1[0][2]
                                                                 bidirectional_1[0][4]
____________________________________________________________________________________________________
lstm_2 (LSTM)                   [(None, None, 300),  541200      embedding_1[0][0]
____________________________________________________________________________________________________
attention_layer (AttentionLayer ((None, None, 600),  540600      bidirectional_1[0][0]
                                                                 lstm_2[0][0]
____________________________________________________________________________________________________
concat_layer (Concatenate)      (None, None, 900)    0           lstm_2[0][0]
                                                                 attention_layer[0][0]
____________________________________________________________________________________________________
time_distributed (TimeDistribut (None, None, 18048)  16261248    concat_layer[0][0]
====================================================================================================
Total params: 27,999,498
Trainable params: 27,999,498
Non-trainable params: 0
____________________________________________________________________________________________________
```

Fig. 1 Hyperparameters of the final attention-based stacked bi-LSTM model

Understanding Call Motivators

We have approached building a classifier in 2 ways, multi-class classification and 1-vs-all classification. Multi-class approach had not resulted in a good model, mainly due to the imbalanced dataset. So, we had dropped it at the early stage of this exercise after trying techniques like over-sampling and under-sampling. We need to build models for 12 motivators as mentioned in the earlier section. We have designed a framework that consists of a variety of techniques at each step and finally selects a model based on the chosen evaluation metric. The process followed to get to 1-vs-all classification models has 5 components as described below:

1. **Input Features**: Count and term frequency inverse document frequency (TF-IDF)-based vectorizers on unigrams and unigrams-bigrams.
2. **Feature Reduction**: PCA (100 dimensions), LDA (100 topics).
3. **Classification Algorithms**: Regularized logistic regressions, support vector machine [28], gradient boosting machines (GBM) [9].
4. **Hyperparameter Tuning**: Using hyperparameter search library Optuna[2] and grid search.
5. **Model selection**: Based on high-performance evaluation metric.

We have applied the same process to the different data sources and combined some of them. The final data sources we have used:

1. **Call transcript**: We have built models separately using transcripts.
2. **Whisper and Summary**: Whisper is stated by the customer and summary (build using repnotes) is from the call transcripts. These 2 datasets complement each other, so it is logical to combine them. These resulted in a good performance on some of the call motivators.
3. **Clickstream data**: We have built models using custom features like, session length, page tags, dwell time, etc.
4. After getting models from each of these datasets, the scores are used as features to build ensemble models for each call motivator. The ensembles have given good lift in precision as well as recall.

The best models for each motivator had different components, some models are built using TF-IDF followed by PCA as input features to logistic regression, some had binary counts followed by LDA as input features to SVM. The final model for 8 of the 12 motivators is an ensemble model built on probability scores from different models. The rest of the models is call transcript-based model with different combination of the above techniques.

[2] https://optuna.org/ (Accessed on 15th June 2021).

5 Experimental Setup and Results

In this section, we narrate the experiments we performed and their results. We performed these experimentations on a Nvidia DGX GPU cluster (having 160,000 CUDA Cores and over 20,000 Tensor Cores) and an internal cloud instance (having 64 GB RAM and 16 processing units).

Experiments related to Call Reason Models

We first tried a simple sequence-to-sequence (s2s) architecture with LSTM. The simple s2s model did not capture the long-term dependencies very well, and since the call transcripts were too long, it did not do well for this task. Next, we tried the s2s model with attention which significantly helped in improving the model's performance (i.e. in minimizing the categorical cross-entropy loss) in the validation set. We experimented with the number of layers in LSTM. We tried 1, 2 and 3 layers. We noticed an improvement in performance while using two layers compared to one layer, but there was no significant performance gain when we tried 3 layers. We also experimented with both unidirectional and bidirectional LSTMs, where bidirectional LSTMs performed better. We were running into exploding gradient issue and applied gradient clipping to avoid it.

We also tried pre-trained transformer-based fine-tuning using BART [13] and T5 [21] but they surprisingly did not do very well. On closer look, it is due to the poorly transcribed data, which did not have well-formed sentences and phrases. Looking at the performance of the above models, we used s2s with the attention model for generating the interaction intent or repnotes.

Model Performance. Since manually validating predictions for the whole data is difficult, we randomly sample 2000 hold-out instances consisting of themes and their corresponding repnotes for evaluating the performance of the model. We also evaluate them automatically. For automatic evaluations we used Rogue-1 and Rogue-L [14]. These results have been presented in Table 3. It reveals that stacked bidirectional LSTM gives the best precision in terms of Rouge score [14]. For manual evaluation, we used the tagged instances and compared them with their corresponding transcripts. We found that in 76.70% of cases, the outputs were acceptable.

Table 3 Summarization model results with highest precisions marked in bold

	Rouge-1		Rouge-L	
Model	Precision	Recall	Precision	Recall
LSTM	0.32	0.16	0.29	0.15
Bi-LSTM	0.36	0.17	0.33	0.16
Stacked bi-LSTM	**0.41**	0.13	**0.40**	0.13
BART	0.38	0.14	0.36	0.14
T5	0.38	0.16	0.37	0.15

Table 4 Model ready pipeline

Module	Techniques	Repnotes
Text pre-processing	Stop word removal, lemmatization, stemming	No
Feature extraction	Count, TF-IDF, BERT embeddings	Yes
Dimensionality reduction	PCA, LDA (Topic models)	Yes

Experimentation related to Call Motivator Models

Due to the nature of call motivators, a one size fits all technique did not work. Hence, we have experimented with a variety of traditional text classification techniques as mentioned in Table 4.

Evaluation Metric. From our experience, we know it is always better to define the metric based on usage of the model by the business. In our case, the predicted motivations for the call will drive the customer-centric solutions. This downstream process becomes costly as it will involve a lot of human resources and capital. If the model is not precise enough, it will lead towards a different solution which might not solve the customer's problem and all this effort might be wasted.

This helped us to decide precision as the metric for evaluation as the cost associated with the wrong prediction is very high. We also ensure that the recall is over a given threshold. Also, we have enough call data to compensate for the loss of recall. The results on the validation set have been presented in Table 5 (P and R denote precision and recall, respectively).

Experimentation with Call Transcript Data (CT). Call transcripts are the most important data stream for predicting the call motivator as it directly comes from the source of the call (call audio). But transcripts can be noisy at times either because of the performance limitation of the speech-to-text system or due to the background

Table 5 Precision (P) and recall (R) on the validation set

Models	CT		WSR		CS		EM	
Motivators	P	R	P	R	P	R	P	R
M1	0.70	0.27	0.57	0.31	0.39	0.77	**0.86**	**0.28**
M2	**0.07**	**0.62**	0.03	0.45	0.05	0.18	**0.07**	**0.62**
M3	0.19	0.33	0.13	0.18	0.04	0.09	**0.88**	**0.25**
M4	0.07	0.46	0.43	0.23	0.00	0.00	**0.43**	**0.23**
M5	0.42	0.31	0.08	0.66	0.16	0.18	**0.92**	**0.25**
M6	0.83	0.15	0.24	0.61	0.26	0.28	**0.83**	**0.19**
M7	**0.86**	**0.14**	0.86	0.14	0.40	0.71	**0.86**	**0.14**
M8	0.66	0.15	0.48	0.08	0.35	0.31	**0.78**	**0.28**
M9	**0.29**	**0.27**	0.10	0.31	0.07	0.20	**0.29**	**0.27**
M10	0.32	0.26	0.29	0.06	0.19	0.18	**0.77**	**0.16**
M11	0.59	0.11	0.15	0.11	0.20	0.22	**0.58**	**0.62**

noise at the side of customer. This directly impacts the performance of the models built on transcripts. After text pre-processing, we have obtained count and TF-IDF vectorizer and used them as features for the model. We also experimented with logistic regression, SVM, GBM and AdaBoost [23]. These models had very high variance, mainly due to the very high number of features. Alternatively, we have used PCA and LDA (which were trained on 1 year worth of call transcripts) to reduce the feature space and later helps us capturing the semantic similarity. These updated models had a comparable performance on the train and validation sets.

Experimentation with Whisper, Call Summaries and Repnotes Data (WSR). Through whisper, we capture direct customer intent and by adding summary/repnotes with it we can capture what has been discussed in the call. For example, a customer may state his need as "*pin password reset*" and summary data has "*customer tried to pin/password reset but the account was blocked*". Both gave us good predictors for "*Attempted to do something online but was unable*". We have observed a good lift on some of the motivators compared to transcript-based models.

Experimentation with Clickstream Data (CS). We have also extracted users' digital footprint of the clicks they had on Fidelity Investment's website. Again, using the web click stream is strongly motivated by our hypothesis that some customers call our customer care if they are themselves unable to do what they want on the website. So, it is fair to assume that we can find signals of call motivation from the click footprint they have left on the website right before the call. For example, it could be as complex as them having tried and failed to place a trade or as simple as them failing in changing their password. As we see here, there could be a spectrum of digital difficulty that could lead a customer to place a call to fidelity's customer care. We are motivated to identify the motivators to push at least these digitally easy tasks towards self-service through the site. From our clickstream data, we extracted user sessions which constituted the sequence of user clicks. We also extracted some other click features such as dwell time, session length, product information of the pages the customer has viewed and user agent attributes such as device information, OS information and browser information. These features had been extracted for various time windows of different lengths. Based on this, we chose 3-time windows (1 day and 1 h before the call and 1 h after the call) for which we pulled all the below features which include both features from the source and engineered features. The list of features used from the source includes—total dwell time, page names, purposes which the web pages are about, products listed in them, channel from which users have come and search phrases. The list of engineered features includes—pre/post/during session times around the call, average of all session lengths of the call, average number of clicks in the sessions of the call, average number of repetitive click instances in the session, the maximum number of repetitive clicks in the session, number of articles viewed and attribute derived from users (such as device, browser and OS), whether virtual agent/chat has been used and whether any page error has been encountered.

Creating ensemble from independent data-based models (EM). After error analysis on predictions of these models, we have observed that these models have some diverse predictions. For the same motivator, where one model failed, other model(s) were driving the decision. These led us towards building ensemble models which take in the predicted probabilities from the individual models as features. We have built logistic regression models for each of the call motivators, which resulted in a good performance on most of the call motivators.

6 Discussion

Analysing the results of the summarization model, we see that its performance in terms of Rouge scores is not as good as that done using manual validation. This is as per the prior works [17] and [20]. To comprehend this, we inspected some instances manually, where the Rouge score is much lower than manual evaluation scores. We found that this happened as the themes (generated using abstractive summarization model) were not syntactically similar to the repnotes but their meanings were similar. For example: "*confirmed beneficiaries on file*" and "*bene verified*". Moreover, when we further analysed the corresponding call transcripts, we observed that several calls had multiple themes in them. However, the repnotes did not cover all of these themes.

With the small data available for detecting motivation of the call, we have experimented with a variety of techniques for feature creation, reduction and classifications using a variety of data sources call transcripts, whisper, clickstream, etc. Call transcript-based models have shown good performance over other independent data sources, except for 1 motivator. The ensemble models have outperformed the independent models, as the model was able to capture information from different sources, e.g. for the motivator "Attempted to do something online but was unable", other than transcript features, the clickstream features like "customer sessions on website" were a powerful signal which has increased precision by 30%. The ensemble model produced around 4 times increase in precision of motivator M3. Similarly, we have seen significant improvement in precision for other motivators. In our next set of experiments, these high precision models can be used to expand the dataset organically with the help of human annotators and help to improve the recall of the models.

7 Conclusion and Future Works

In this paper, we have introduced an approach that synthesizes the reason for the call with the motivation behind it. It not only extracts call reasons but also determines what the primary motivator was behind the call. Presently, this is being put into production to generate insights that are more informative and interpretable by the business. It drives them to take action by strategizing these call reasons and motivators as internal guides to drive business decisions, reduce the calls to action, drive business digitally and boost the performance of specific key performance indicators significantly. The results confirm the benefits of our approach. For further work, the possibility of learning with other structured outputs, increasing the tagged data using a semi-supervised graph algorithm is recommended. Moreover, we would like to extract multiple themes from a call using beam search like approach. We would like to monitor the performances of these models and evaluate them using external factors like reduction in call volumes, increment in customer experience index and so on.

References

1. Agrawal Y, Anand V, Arunachalam S, Varma V (2021) Hierarchical model for goal guided summarization of annual financial reports. In: Companion proceedings of the web conference 2021, pp 247–254. WWW '21, Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3442442.3451373
2. Ait Azzi A, Kang J (2020) Extractive summarization system for annual reports. In: Proceedings of the 1st joint workshop on financial narrative processing and multiling financial summarization, pp 143–147. COLING, Barcelona, Spain (Online). https://www.aclweb.org/anthology/2020.fnp-1.24
3. Chopra A, Nagwanshi P, Ghosh S (2021) Data driven content creation using statistical and natural language processing techniques for financial domain. arXiv preprint arXiv:2109.02935
4. Chu-Carroll J, Carpenter B (1999) Vector-based natural language call routing. Comput Linguist 25(3):361–388. https://www.aclweb.org/anthology/J99-3003
5. Ciravegna F, Gilardoni L, Lavelli A, Mazza S, Black WJ, Ferraro M, Mana N, Matiasek J, Rinaldi F (2000) Flexible text classification for financial applications: the facile system. In: Proceedings of the 14th European conference on artificial intelligence, pp 696–700. ECAI'00, IOS Press, NLD
6. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
7. El-Haj M, AbuRa'ed A, Litvak M, Pittaras N, Giannakopoulos G (2020) The financial narrative summarisation shared task (FNS 2020). In: Proceedings of the 1st joint workshop on financial narrative processing and multiling financial summarization, pp 1–12. COLING, Barcelona, Spain (Online). https://www.aclweb.org/anthology/2020.fnp-1.1
8. Erkan G, Radev DR (2004) Lexrank: graph-based lexical centrality as salience in text summarization. J Artif Int Res 22(1):457–479
9. Friedman J (2001) Greedy function approximation: a gradient boosting machine. Annal Statist 29:1189–1232. https://doi.org/10.2307/2699986
10. Gupta V, Lehal GS (2010) A survey of text summarization extractive techniques. J Emerg Technol Web Intell 2:258–268
11. Higashinaka R, Minami Y, Nishikawa H, Dohsaka K, Meguro T, Kobashikawa S, Masataki H, Yoshioka O, Takahashi S, Kikui G (2010) Improving hmm-based extractive summarization for multi-domain contact center dialogues. In: 2010 IEEE spoken language technology workshop, pp 61–66. https://doi.org/10.1109/SLT.2010.5700823
12. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735, https://doi.org/10.1162/neco.1997.9.8.1735
13. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V, Zettlemoyer L (2020) BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 7871–7880. Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2020.acl-main.703, https://www.aclweb.org/anthology/2020.acl-main.703
14. Lin CY (2004) ROUGE: a package for automatic evaluation of summaries. In: Text summarization branches out, pp 74–81. Association for Computational Linguistics, Barcelona, Spain, https://www.aclweb.org/anthology/W04-1013

15. Litvak M, Last M, Friedman M (2010) A new approach to improving multilingual summarization using a genetic algorithm. In: Proceedings of the 48th annual meeting of the association for computational linguistics, pp 927–936. Association for Computational Linguistics, Uppsala, Sweden, https://www.aclweb.org/anthology/P10-1095
16. Litvak M, Vanetik N (2013) Mining the gaps: towards polynomial summarization. In: Proceedings of the sixth international joint conference on natural language processing, pp 655–660. Asian Federation of Natural Language Processing, Nagoya, Japan, https://www.aclweb.org/anthology/I13-1075
17. Liu F, Liu Y (2008) Correlation between ROUGE and human evaluation of extractive meeting summaries. In: Proceedings of ACL-08: HLT, short papers, pp 201–204. Association for Computational Linguistics, Columbus, Ohio, https://www.aclweb.org/anthology/P08-2051
18. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: a robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692
19. Mihalcea R, Tarau P (2004) TextRank: bringing order into text. In: Proceedings of the 2004 conference on empirical methods in natural language processing, pp 404–411. Association for Computational Linguistics, Barcelona, Spain, https://www.aclweb.org/anthology/W04-3252
20. Ng JP, Abrecht V (2015) Better summarization evaluation with word embeddings for ROUGE. In: Proceedings of the 2015 conference on empirical methods in natural language processing, pp 1925–1930. Association for Computational Linguistics, Lisbon, Portugal. https://doi.org/10.18653/v1/D15-1222, https://aclanthology.org/D15-1222
21. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ (2020) Exploring the limits of transfer learning with a unified textto-text transformer. J Machine Learn Res 21(140):1–67, http://jmlr.org/papers/v21/20-074.html
22. Reimers N, Gurevych I, Reimers N, Gurevych I, Thakur N, Reimers N, Daxenberger J, Gurevych I (2019) Sentence-bert: sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 conference on empirical methods in natural language processing. Association for Computational Linguistics
23. Schapire RE (1999) A brief introduction to boosting. In: Proceedings of the 16th international joint conference on artificial intelligence, Vol 2, pp 1401–1406. IJCAI'99, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA
24. See A, Liu PJ, Manning CD (2017) Get to the point: summarization with pointergenerator networks. In: Proceedings of the 55th annual meeting of the association for computational linguistics, Vol 1, Long Papers, pp 1073–1083. Association for Computational Linguistics, Vancouver, Canada. https://doi.org/10.18653/v1/P17-1099, https://www.aclweb.org/anthology/P171099
25. Shi T, Keneshloo Y, Ramakrishnan N, Reddy CK (2021) Neural abstractive text summarization with sequence-to-sequence models. ACM/IMS Trans Data Sci 2(1). https://doi.org/10.1145/3419106, https://doi.org/10.1145/3419106
26. Singh A (2020) PoinT-5: pointer network and T-5 based financial narrative summarisation. In: Proceedings of the 1st joint workshop on financial narrative processing and multiling financial summarization, pp 105–111. COLING, Barcelona, Spain (Online), https://www.aclweb.org/anthology/2020.fnp-1.18
27. Sun F, Belatreche A, Coleman S, McGinnity TM, Li Y (2014) Preprocessing online financial text for sentiment classification: a natural language processing approach. In: 2014 IEEE conference on computational intelligence for financial engineering economics (CIFEr), pp 122–129. https://doi.org/10.1109/CIFEr.2014.6924063
28. Vapnik V, Guyon I, Hastie T (1995) Support vector machines. Mach Learn 20(3):273–297
29. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser LU, Polosukhin I (2017) Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems, Vol 30. Curran Associates, Inc., https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

30. Verma R, Chen P, Lu W (2007) A semantic free-text summarization system using ontology knowledge. In: Proceedings of the document understanding conference
31. Yang L, Kenny E, Ng TLJ, Yang Y, Smyth B, Dong R (2020) Generating plausible counterfactual explanations for deep transformers in financial text classification. In: Proceedings of the 28th international conference on computational linguistics, pp 6150–6160. International Committee on Computational Linguistics, Barcelona, Spain (Online). https://doi.org/10.18653/v1/2020.colingmain.541, https://aclanthology.org/2020.coling-main.541
32. Zhao W, Zhang G, Yuan G, Liu J, Shan H, Zhang S (2020) The study on the text classification for financial news based on partial information. IEEE Access 8:100426–100437. https://doi.org/10.1109/ACCESS.2020.2997969
33. Zheng S, Lu A, Cardie C (2020) SUMSUM@FNS-2020 shared task. In: Proceedings of the 1st joint workshop on financial narrative processing and multiling financial summarization, pp 148–152. COLING, Barcelona, Spain (Online), https://www.aclweb.org/anthology/2020.fnp-1.25

Prediction of Indian Currency for Visually Impaired People Using Machine Learning

Arpan Maity, Srijita Guha Roy, Sucheta Bhattacharjee, Priyanka Dutta, and Jagannath Singh

Abstract Most of the time the people are finding it difficult to recognize the paper currencies due to their inability to see. Also in real life, it is not possible to trust the people to help a visual impaired person in recognizing the paper currency. But, in this regard, we can always trust an application that is easy to access and trust worthy. For building the application for the same, we need the help of machine learning algorithms. Several algorithms of machine learning such as supervised, unsupervised, semi-supervised, and reinforcement learning are available. Our paper mainly focuses on the recognition of different currency notes for the visually impaired people by using machine learning and taking the model as ResNet50 which is a 50 layers deep convolutional neural network. We have also used image processing by resizing and cropping the image to extract the important features of the notes so that the machine can recognize the currency notes efficiently. Our main contribution in this paper is through using various methods of image processing and feeding those processed image in our model, so that this model gets a fairly high accuracy.

Keywords Currency detection · Image processing · Image cropping · Convolutional neural network (CNN) · Preprocessing

1 Introduction

Currency is a medium of exchange for services and goods which is found in the form of papers or coins and is issued by the government of a country. Paper currency is a commonly used as a medium of transaction because of its durability and convenience. Currently, the notes are the most preferred form of currency compared to its other alternatives. But, it becomes difficult for visually impaired people to recognize a paper currency. 2.2 billion people are found and classified to be visually impaired

A. Maity (✉) · S. G. Roy · S. Bhattacharjee · P. Dutta · J. Singh
KIIT Deemed to be University, Bhubaneswar, India
e-mail: maityarpan797@gmail.com

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_19

worldwide (according to WHO)[1], out of which 37 million [1] people belong to India. Out of many, the most common main problem faced by people with visual impairment is the incapacity to identify the paper currencies because the similar paper texture and size between the different currencies [2]. Between the old 50, 100, 500, and 1000 rupee banknotes, there was a discernible size difference of 10mm or more either by length or width, a distinction that helped most of India's blind population transact with cash. But in the fall of 2016, the Indian government banned 500 rupee and 1000 rupee bills[2] and introduced higher denomination banknotes in the market. For India's visually impaired population, the move affected their mobility. Hence, with the introduction of new currency notes into the Indian economy, recognition of currency became a necessity for the visually impaired people as they are unable to differentiate between currencies properly. As a result of this, they can be cheated by others easily.

This paper mainly offers a machine learning (ML)-based model to help visually impaired people. In this model, we have used *image processing* [3] along with *image cropping* [4] to extract essential features of an image so as to process in an efficient way using ML that will allow visually impaired people to identify different currency notes. Therefore, our main contribution by proposing this model is to provide an easy solution to resolve this trouble so that the visually impaired people feel safe and confident in their financial approach.

Our main contribution can be summarized as follows

- We have collected the datasets from various sources using different cameras in different lighting conditions
- Not only we had taken only flat and front side of the notes, we have also taken backside of the notes and half folded of notes, do replicate the real-life scenarios
- Furthermore, we had preprocessed our dataset so that the accuracy of the predictions can increase.

The rest of the paper is structured as follows: In Sect. 2, we presented the basic concepts of machine learning, image processing, and image cropping and resize. In Sect. 3, we present the existing related work. Section 4 presents about the dataset collection and processing. Sect. 5 contains the details of the experimental setup, our ML model, building that model and accuracy, and Sect. 6 contains future perspectives of the work and concludes the paper.

2 Basic Concepts

In this section, some basic concepts are presented, which are very relevant and important for understanding our proposed technique.

[1] https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment.

[2] https://m.economictimes.com/news/politics-and-nation/in-an-attempt-to-curb-black-money-pm-narendra-modi-declares-rs-500-1000-notes-to-be-void-from-midnight/articleshow/55315932.cms.

2.1 Machine Learning

Machine learning is an application of artificial intelligence (AI) [5] that provides systems the flexibility to learn automatically on their own, improve and predict results from their prior experience without being programmed explicitly. The learning process starts with the gathering of knowledge, analysis of information through certain instructions, so as to make better decisions and form and predict patterns in data through the provided examples through the sample datasets. The first aim is to make the computers learn all by itself without any human interference or support and can adjust their actions according to the information they learn. The focus of the sector is learning, that is, acquiring skills or knowledge from experience. Most ordinarily, this suggests synthesizing useful concepts from historical data. As such, there are many alternative models of learning that we may encounter as a practitioner within the field of machine learning, like regression, which is employed to identify the connection between a dependent variable quantity and one or more than one independent variables and is usually leveraged to form predictions about future out comes. When there is only one variable and one variable quantity, it is referred to as simple statistical regression. Because the number of independent variables increases, it is said as multiple statistical regression. The model utilized in this project is the ResNet50(f) which is largely a deep convolutional neural network [6], details given in Sect. 6.2 (Fig. 1).

Artificial neural network (ANN) [7] is programmed like the brain of a human, which contains thousands of neuron, network of nerves interconnected sort of a web. The processing units of an ANN are made of input and output units and has hundreds or thousands of artificial neurons, and they are actually connected with each other by nodes. Various structures and forms of data supported an inside weighting system are received by input units.

Fig. 1 Artificial neural network

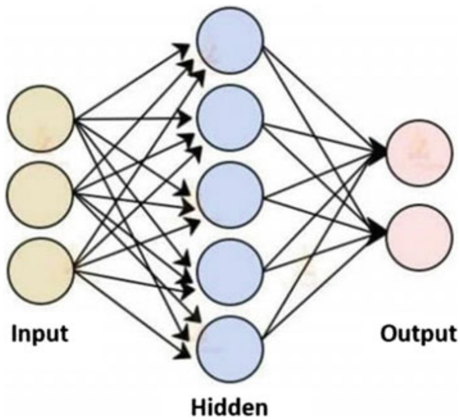


Fig. 2 Example of image processing

2.2 Image Processing

Image processing [3] is a technique with which certain operations are performed on a photo, in order to convert it into an advanced image or to find out some meaningful or important information from it for coding or other purposes. It is the type of processing within which both the input and the output is a picture. At times, image processing is also done to enhance the standard of the image being given as an input. Just like human brains, machines can also be trained to interpret images to research those images rather even more effectively than we will. After applying image processing, artificial intelligence (AI) can ensure security by using power face recognition and authentication functionality. This will ensure security publicly places [8], recognizing and detecting patterns and objects in images and videos, and so on. Several advanced image processing methods form machine learning models like deep neural networks to rework the pictures on a spread of tasks, like applying filters which are artistic, tuning a picture for best quality, or increasing certain image parameters, like contrast, saturation level, etc., to maximize the image quality for computer vision programs (Fig. 2).

2.3 Image Cropping and Resize

Image cropping and resize image cropping [4] may be a common photo manipulation process, which improves the overall composition by removing unwanted regions. Image cropping is widely used in graphic design, photographic, film processing, and printing businesses. To "crop" an image is to get rid of or adjust the skin edges of a picture to enhance framing or composition, draw a viewer's eye to the image subject, or change the dimensions or aspect ratio. In sort, image cropping is nothing but

Fig. 3 Example of image cropping

removing or clipping out unnecessary or unimportant part of the images, for better processing (Fig. 3).

Resizing is altering the scale of your image without cutting anything out. To resize a picture is to essentially change the file size. In the vision of computers, resizing images could be a critical preprocessing step. Mostly, with images of smaller size, the training of our machine learning models is faster. Moreover, many ML programmers require that the photos are of the same resolution and size because the raw images gathered may differ in size. So as to beat this flaw, we resize the photographs so machine can predict efficient results from the image dataset provided because the input.

3 Related Works

In this section, we review some of the work related to the currency identification. In our literature survey, we found that many work has been done in the field of currency prediction using image processing. We have presented few of the closely related works along with their limitations.

In a work by Zhang and Yan [9], they have used single shot multi box detector (SSD) model which is based on the concepts of deep learning, and the convolution neural network (CNN) model to list the properties of notes from both front and back sides, so that the values of the currency can be predicted accurately. Single shot multi box detector (SSD) is a big frame model for detection of objects. It uses feature maps on different layers to generate bounding boxes, and the output of these bounding boxes is obtained. In the process of detection, after the pixel problem, the SSD eliminates the requirement for any further extraction of features. The SSD is easy to train as in each and every layer it stores the calculated results of different stages.

In this paper [9], they have used three different New Zealand dollars (NZD), i.e., 5NZD, 10NZD, and 20NZD for currency recognition. After the dataset was collected, data augmentation was done to increase the size of the dataset. Initially, there were 300 pictures but after taking pictures from different heights, angles, and positions the enriched dataset contained 7500 pictures. The resolution of each picture was 1280×720. The collected data was then given for data training and validation, and the overall learning framework was TensorFlow. For the currency recognition, the accuracy was based on the accuracy of training as well the verification dataset.

All the existing solutions [10] of the currency recognition are very costly and require special hardware. To improve the predictions, segmentation method has been followed here. By segmentation, the unwanted background was possible to be separated from the foreground. To do that, and to detect the appropriate region of interest (ROI), grabcut algorithm was used. A visually impaired person will find it difficult to identify which one is the front side of the currency and which one is the back side. That's why a commendable number of images in the dataset was backside of the note. Images for half folded notes were also included for better accuracy. By doing all these, the accuracy of the dataset reached to 96% which is 4.5% better when executed without segmentation method, when tested with 391 test images. To improve the performance of the application, all the dataset images were at first compressed, so that they take minimum space, then they were stored locally in the disk, so that the time complexity decreases. Storing the files in cloud, would decrease the app size to a certain extent, in the cost of higher processing time, provided that the user has a great internet connectivity, which is not favorable. By doing, his average processing time was reduced to 1.2 s.

Another work by Kamesh et al. [11] proposed an ingenious model that helps the visually impaired people for using currency. The bank notes are scanned with the help of various image processing methods, the noise is removed, and the region of interest (ROI) is marked. Then, the image is converted to text format and from the text to audio which can be clearly heard by the visually impaired people. The proposed model implements the Raspberry Pi micro controller-based system. An IPR sensor is being used in the prototype model for the object detection instead of a camera. An effective method has been used here which helps to identify the object, mark the ROI within the object, scan the text objects and finally generate binary characters from the scanned text through optical recognition. Then, different currency notes have been tested by the embedded board and the speech is generated in English to identify the value of the currency. In the future, various other national and international languages can also be used to generate the speech. This system also helps the visually challenged or impaired people in their day to day activities like reading, by converting texts into an audio format. The detection of obstacles can also be carried out using PIR sensor, and the currency value can also be found out with the help of a database. The PIR sensor cannot name the object, but it alerts the visually impaired user about the obstacle.

The currency detection using similarity indices method proposed by Vishnu et al. [12], and a measurement of similarity was used for classifying the paper currency on the basis of similarity of the features extracted. A currency has two kind of features one is the extrinsic feature which is the physical feature that includes the currency size, whereas the other one is the intrinsic feature which includes texture and color of the currency. This method uses the color and the shape to recognize the value of the currency and the validity check from the recognition made by the color and shape feature is ensured by the other four features. To prepare the currency dataset, both sides of a currency were scanned with a 150 dpi resolution color scanner and the auto cropping was used while scanning to get the exact size of the currency. The region of interest (ROI) was fixed at particular locations on the image of the currency. The ROI includes the color information, shape ROI, center ROI, Ashoka, RBI seal, and the RBI Governor's signature. The area that comes within the ROI was cropped and then saved as a different images for the recognition process. The similarity value between all the currencies having the same features in the dataset was calculated. After that, the average value of similarity was also computed and that feature of the currency having the highest average similarity was chosen as a template feature for testing the other currencies. This process was conducted on the scanned images of a currency dataset including values of 50, 100, 500, and 1000. Then, six features from all currencies in the database were extracted. The information on color helps to select the suitable template feature for a particular currency, and the shape was determined with the help of the shape detection method. The RBI seal present in the currency was determined by recognizing which color is more dominant. The remaining three features were determined by evaluating the similarity with the features of the test samples. This proposed model works on a dataset of 200 scanned images of currencies having different values and has an accuracy of 97%.

4 Dataset Collection and Processing

In our project work, the data is taken from Mendeley Data version 1 and two different datasets from Kaggle. Published in August 2020, Mendeley Kaggle[3] keeps a huge variety of datasets for new Indian currencies. Dataset is created using a 21 MP camera. All the currency images in the dataset were captured in landscape mode, and their size is 5344×3000. Whereas the size of all the currency images in the dataset which are captured in portrait mode is 3006×5300. Kaggle[4] is also a renowned website created in 2010 created by a team of data scientists. It helps users to find, publish, and work on datasets, explore and build models in a web-based data-science environment, work with some other ML professionals and data scientists, and to solve challenges based on data science in different competitions. Our dataset contains a total of 1085 images, where 329 were taken from Mendeley and the rest is taken from Kaggle. We included both the new and old notes of Rs. 10, 20, 50, and 100 to deal with every possible scenarios.

[3] https://data.mendeley.com/datasets/48ympv8jjf/1.

[4] https://www.kaggle.com/gauravsahani/indian-currency-notes-classifier.
https://www.kaggle.com/vishalmane109/indian-currency-note-images-dataset-2020.

4.1 Dataset Optimization and Preprocessing

We needed to image processing algorithms in this project. But the main problem of that was as follows:

1. Photos were of various shapes and sizes. To implement the same algorithm for all the images, we need to get the all the photos into same shape and size. Hundreds of photos having different sizes could increase the overall complexity of our program. So we analyzed all the images to find an optimal size, so that all the images can be resized, without much distortion. We chose 1300×1000 as the optimal size and wrote an algorithm so that all the images can be resized at once
2. Second most important problem was the object to frame ratio. As we selected our dataset from different sources, so the object to frame ratio varied vastly. In some images, the notes occupied about 90% of the frame, and in some other, the object (note) occupied as low as 35% of the total frame. Though this scenario will not create any errors, but will adversely affect the accuracy of our model. So we needed to improve that. We observed all the images closely to get the optimal coordinates, so that each and every image can be cropped using the same values, without much data loss. We chose (239, 200) as the top-left coordinate and (1200, 800) as the bottom-right. As a result, our average object to image ratio increased from about 63% to 86%. We wrote an algorithm to perform the cropping of all the images all at once and the time taking for doing all of this was 11.92 s, which is considered quite good.

5 Algorithm Used and It's Flowchart

The step by step method, that we followed, for this program is discussed below (Fig. 4)

- **Resizing**: First, each and every image is resized to an optimal size. The reason behind it was that, since, all the images had different origins, they had different sizes, so to apply our ML algorithm to it, we applied resizing
- **Region Detection**: After that, to get only the useful part of the image, we applied region detection to it. Detail description is in Sect. 4.2
- **CNN Processing**: After that, we used deep spatial feature extraction using convolutional neural network (CNN)
- **Feature extraction**: Parallelly with that, we also did an extensive feature extraction using residual network version 2 model
- **Feature fusion**: Combining the above two methods, we derived all the features of our classifier model. Details about each of the layers of our CNN model using ResNetV2[5] is provided in Diagram 4 (b)

[5] https://subscription.packtpub.com/book/big-data-and-business-intelligence/9781788629416/2/ch02lvl1sec13/resnet-v2.

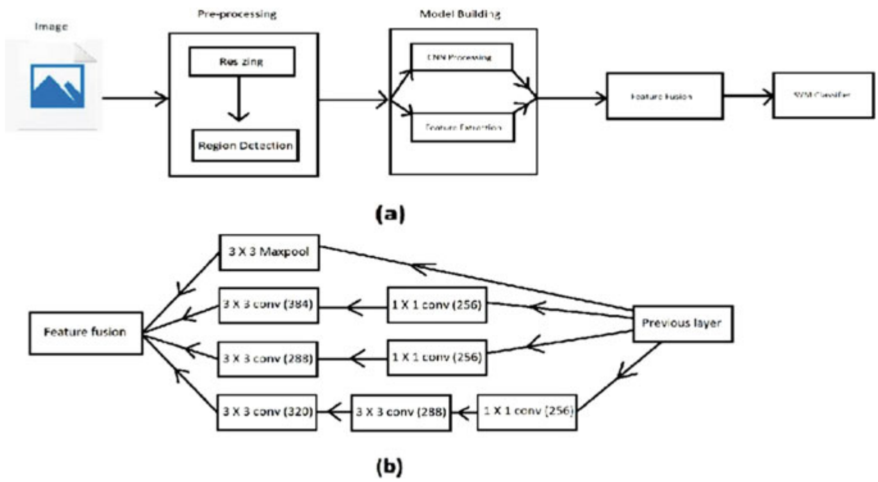


Fig. 4 Flowchart of algorithm

- **SVM Classifier:** After this, the whole model is fed into SVM classifier, for better efficiency and accuracy

6 Experimental Setup and Methodology

6.1 Experimental Setup

All the member who participated in this work used the same device. We used a HP ProBook laptop with Intel(R) Core(TM) i5-8250U processor, CPU @ 1.60, and 1.80 GHz. RAM available was 16 GB. So the processing time may vary a bit from device to device.

In our work, we had used TensorFlow, which is a machine learning platform that allows users to build ML models by providing an open source, in-memory, distributed, ML and predictive analytic. TensorFlow allows developers to visualize and work with medium to large sized datasets. It also helps us analyze how the data moves through a graph by using dataflow graphs, or a series of processing nodes. A mathematical operation is represented by each node in the graph, and each connection or edge between nodes is a multidimensional data array or tensor.

	Rs. 10	Rs. 20	Rs. 50	Rs. 100	Rs. 200	Rs. 500	Rs. 2000
Rs 10	32	1	0	0	0	1	0
Rs 20	0	30	0	1	0	0	0
Rs. 50	0	0	29	0	1	0	0
Rs. 100	0	1	0	35	0	0	0
Rs. 200	0	0	0	2	30	0	0
Rs. 500	0	0	0	0	0	36	1
Rs. 2000	0	0	0	0	0	0	21

Fig. 5 Confusion matrix

6.2 Choice of Machine Learning Algorithm

In this project, we chose ResNet50 (full form residual network 50) as our ML algorithm. ResNet50 has 48 convolution layers along with 1 MaxPool and 1 average pool layer and is also a variant of ResNet model. It consists of 3.8×109 floating points operations. It was our go-to model for currency detection as it allowed to train extremely deep neural networks with 150+ layers with a great a level of accuracy. This can be verified with the help of the confusion matrix in Fig 5.

Hence, the choice of ResNet50 as our go-to ML algorithm due to the low execution time and high accuracy is proved right.

6.3 Class Imbalance Problem

While choosing the datasets, we kept a keen eye on equality of all the datasets, so that we can prevent the class imbalance problem. By doing so we kept 155–160 image in each class (in total there were seven classes) and successfully prevented the problem

6.4 Model Building

We split our dataset into train, test, and validate in the ratio of whole numbers of 76: 12: 12. We used flow_from_directory() method so that all the images from the dataset can be fed to our model at once. We introduced 50 epochs for gaining better accuracy. We executed our code several times before and after preprocessing. The result of the continuous change in epoch is demonstrated in Fig. 6.

We know that a continuous horizontal epoch at 1.0 without much deviation is the sign of a good overall accuracy. So we can conclude from this discussion that our

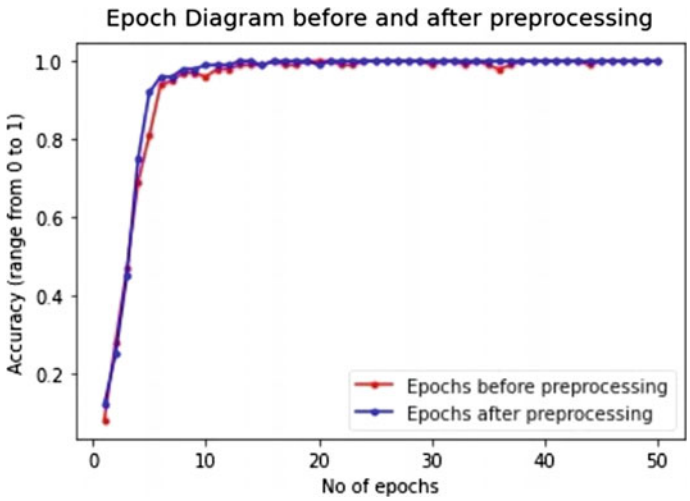


Fig. 6 Accuracy of 50 epochs before and after preprocessing

image preprocessing was effective, which resulted about 8% better accuracy on an average

6.5 Additions

We recognized the fact that since the model is for visually impaired people, so in real time use scenario, it will be difficult for them to detect whether they are holding the notes in the correct way or upside down. They will also we unsure about whether the notes are folded or not. So we included almost 400 of such images which are captured in backside up way or in folded manner, for the betterment of performance of the model in real life use cases.

6.6 Model Accuracy

As discussed in Sect. 6.4, the more the line of epoch[6] is horizontal (or undeviated) in 1.0, the more is the accuracy of the model, which was the exact case when we ran the code after preprocessing. The accuracy which was hovering around 86–88%, became around 95% (97.9% at max) after applying the preprocessing. To confirm the changes, we ran the code 25 times (both before and after preprocessing), whose result is shown in Fig. 7. All of these confirms about our successful model building.

[6] https://radiopaedia.org/articles/epoch-machine-learning.

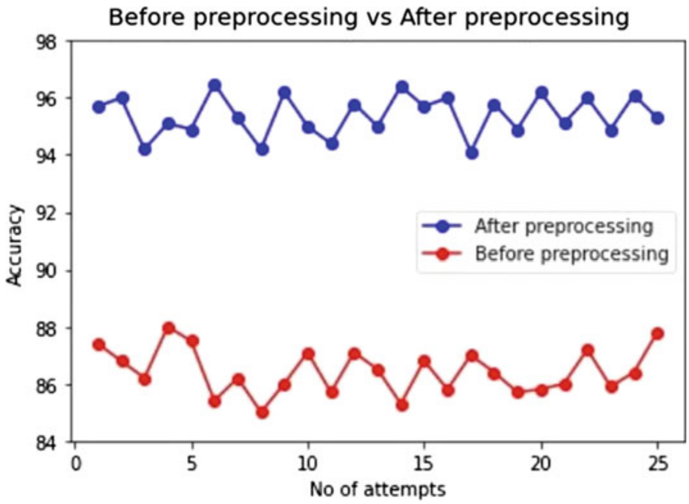


Fig. 7 Accuracy before and after preprocessing for 25 attempts

The total accuracy is calculated with the help of the following equation:

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{False Negative} + \text{True Negative} + \text{False Positive}}$$

This accuracy rate is also confirmed by the confusion matrix, shown in Fig. 4.

7 Conclusion and Limitations

The most important feature of our work is that it is device independent, and people do not need a high resolution camera to capture the images. A device with 4 MP camera or above, is more than enough. The result may also vary based on the lighting conditions. As the image is cropped on, if we crop an image captured in low lighting condition, then there is a possibility that the resultant image becomes blurry which will affect the overall result. Though the position and orientation of the note will not affect the results as discussed in Sect. 5.5. We have implemented the proposed technique of currency recolonization without image cropping and with image cropping. It was observed that the accuracy of the proposed technique is 86% without image cropping and 95% with image cropping.

In the future, we can improve the preprocessing drastically. In this project as discussed in Sect. 4.1, we searched for an optimal size of the image to resize the image, and cropped so that the object to image ratio constant. In later part of our work, we can use object detection method of ResNet50 so that the model becomes yet more accurate. We can also implement some other modern computer vision algorithms to make the predictions more robust.

References

1. Pathak A, Aurelia S (2020) Mobile-based Indian currency detection model for the visually impaired. In: Convergence of ICT and smart devices for emerging applications. Springer, Cham, pp 67–79
2. Patil MV, Vishal MA, Ajay Sirsat (2020) Artificial intelligence digital assistant for visually impaired people. Int J Fut Gen Commun Network 13(1):24–31
3. Mirza R, Nanda V (2012) Paper currency verification system based on characteristic extraction using image processing. Int J Eng Adv Technol (IJEAT) 1(3):68–71
4. Ciocca G et al (2007) Self-adaptive image cropping for small displays. IEEE Trans Consum Electron 53(4):1622–1627
5. Yang, Qiang, et al. "Federated machine learning: Concept and applications." ACM Transactions on Intelligent Systems and Technology (TIST) 10.2 (2019): 1–19.
6. Ghosh D, Singh J (2020) A novel approach of software fault prediction using deep learning technique. In: Automated software engineering: a deep learning-based approach. Springer, Cham, pp 73–91
7. Singh J, Sahoo B (2011) Software effort estimation with different artificial neural network
8. Li L et al (2020) A review of face recognition technology. IEEE Access 8:139110–139120
9. Zhang Q, Yan WQ (2018) Currency detection and recognition based on deep learning. In: 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS). IEEE
10. Chakraborty K et al (2013) Recent developments in paper currency recognition system. Int J Res Eng Technol 2:222–226
11. Kamesh DBK et al (2016) Camera based text to speech conversion, obstacle and currency detection for blind persons. Indian J Sci Technol 9(30):1–5
12. Vishnu R, Bini Omman (2014) Currency detection using similarity indices method. In: International conference for convergence for technology-2014. IEEE

Comparative Analysis of Different Machine Learning Prediction Models for Seasonal Rainfall and Crop Production in Cultivation

S. Deepa and S. Gokila

Abstract Agriculture is one of the strengths of India, from the last few years, gradually the agriculture growth is going downwards in other side the population growth is upwards. Reason for agricultural downward growth depends on so many parameters. The rainfall is one of the main parameters which affects the crop yield. Because of this, the farmers are also facing the loss. If they know this information in prior, the farmers can plan accordingly the type of crop suited for the particular season and it helps the farmer to get good profit out of it. Machine learning scientific and statistical methods are used for predicting the rain fall and crop yield. Kharif and Rabi are two seasons taken for analysis. The regressor predicting models are constructed to predict the seasonal rainfall and crop yield. This study primarily focuses on seasonal crop production prediction, which is dependent on rainfall. The different types of machine learning regression method are used to achieve better results. The performance of comparison models is evaluated using different metrics. Finally, the linear regression and Bayesian linear regression models comparatively produce the best result in terms of accuracy for rainfall prediction. The boosted decision tree regression model is achieving the better result for crop prediction.

Keywords Crop prediction · Regression model · Linear model · Machine learning · Rain fall

1 Introduction

Agriculture is the primary industry in many of the countries. The traditional farming had happened for months to grow crops and cultivate it. In modern years, the crop growing duration has been reduced to a few months. So the land is being utilized for

S. Deepa (✉)
School of Sciences, CHRIST (Deemed to be University), Bengaluru, India
e-mail: sdeepa369@gmail.com

S. Gokila
Department of Computer Applications, Hindustan Institute of Technology and Science, Chennai, India

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_20

an entire year to grow different varieties of crops. In this aspect, the seasonal wise crop prediction system will be a great support system for farmers.

All kinds of food and cash crops are mainly categorized using the soil type and the season. In India, rainy season is playing important role in deciding the share of agriculture in gross domestic product (GDP) of the year [1].

The prediction supports the farmer to decide the seasonal crop, to balance demand and supply and the expert organization to make policy. Identifying the crop suitable for the farmer area will be great support to get more yield and also to fix the price [1]. There are many supporting systems that exist to provide them guidance.

The branch of artificial technology that delivers the ability to study and progress by frequent training is machine learning. The machine learning technology contains well-defined models to college specific data and apply correct algorithms to get accurate results [2]. The ML techniques are applied in various domains, one among the major domains in agriculture. The techniques have been applied in this domain to predict seed, crop, yield, classify soil, etc. The season wise crop production prediction is a great support for the farmers to get maximum yield to manage the demand and supply. The ML algorithms are used in this work to predict the crop production of Himachal Pradesh for Rabi and Kharif season.

Various different models have been projected and validated for crop yield prediction. This prediction depends on various factors like weather, climate, soil, seed and fertilizer. So it uses several data sets [2]. The ML algorithm can determine the patterns and correlation among the data to discover the knowledge from that. The model is trained using past experience data, and the outcome is represented based on that. The model had been built with several factors, and the parameters to the model are identified using historical data during the training phase. Certain ratio of data used for the training phase, and the remaining data have been used to evaluate the model in the testing phase. The ML model can be descriptive in the explanation of predictive occurrence. The machine learning technique can act as a parallel thinker giving the technological contribution in decision making.

The previous production history of various crops along with rainfall data has been taken in this analysis. The cultivation of crops varies based on the geographical area. The region taken for work is rich in water source and soil nature [need to find source about HP]. Due to the climate change and variation in rainfall, the season wise analysis is an apt one when done along with rainfall detail. The season wise split done in proposed work gives more accuracy in production prediction.

2 Literature Review

Five supervised machine learning algorithms namely K-nearest neighbour (KNN), multinomial logistic regression, Naïve Bayes, random forest and artificial neural network (ANN) are applied [2]. The soil is taken as a major feature, and yield of mustard crop was predicted. Among which ANN and KNN were considered to be more precise methods. The work helps to identify the best algorithm.

AdaSVM and AdaNaive Bayes are proposed ensembles to project crop production in prediction models [3]. The methodology applied for five different crops using two different data sets containing weather parameters. The proposed methods are compared with classic SVM and Naïve Bayes also. Among all the methods, AdaNaive Bayes produced high accuracy for all the five crops.

The linear predictive model was applied to predict the yield of crop [4] for farmer location. The method produced 85% of accuracy. SVM and linear regression techniques to predict suitable crops by using rainfall, soil quality and soil composition are applied [5]. The documents generate an idea of trying with nonlinear algorithms, whereas the climate related features are nonlinear in nature [6]. The SVM classification applied to identify the seed which gives high crop production. The accuracy of seed classification is directly proportional to feature extraction. The test data accuracy in SVM average filter with 300 extracted features produced 95.88% accuracy. The correct seed selection to increase the crop production was done with temperature, humidity and moisture content parameters of 10 years of data using Naïve Bayes machine learning technique [7].

The work took soil as a major feature to predict the crop yield in Maharashtra region [8]. This work used KNN, KNN with cross validation, Naïve Bayes and support vector machine algorithms among which KNN with cross validation obtained 88% accuracy.

The rainfall prediction using SVM and crop prediction using decision tree algorithm had been implemented in [9]. The rain prediction took temperature and humidity as required parameters. The rainfall, temperature, humidity and soil pH are taken as features to do the crop prediction. The analysis done for the year 2020. The methodology gives an idea to take more years of data to train ML models [10]. The work integrated data from multiple sources for soil, weather and production details and applied multiple linear regression models for crop prediction.

Three different data sets for soil, rainfall and yield are combined for sugarcane crop yield prediction using supervised learning techniques like KNN, SVM and least square support vector machine [11]. The comparative analysis shows that LSSVM gives accurate predictions while comparing to other two methods. The LSSVM produced 90% accuracy.

The crop yield prediction was done by considering cost of cultivation, cost of irrigation and cost of production as independent variables using decision tree method, KNN and Naïve Bayes classifiers. The result of the work shows that KNN produces higher accuracy than the other two methods. Among the reviewed comparative analysis research the KNN produces higher accuracy of prediction. In this work, also KNN produced 89.4% accuracy [12].

Random forest classifier with soil, min temperature, max temperature humidity and rainfall of India predict millet crop production was applied [13]. The methodology produced 99.74% of accuracy. One of the recurrent network model is long short-term memory. The random forest supervised machine learning algorithm applied on training phase to multitude the decision tree form classes and regression prediction was applied. The method predicted all the crops of a particular district with 75% accuracy [14].

The machine learning and deep learning approaches can be applied in pre-sowing process. It also helps to come out of a regular crop and suggests the best suited crop which gives high production to reach profit [15, 16]. The object of the review work is to exhibit the research carried out in a specific field [1]. Such a review report concludes that RNN with a hybrid algorithm of deep learning produces 90% accuracy in prediction. The hybrid methodology outperforms when compared to the ANN, CNN and RNN-LSTM deep learning methods.

The RNN, LSTM applied for vegetable (carrot, beans, beetroot and onion) prediction, and ARIMA model was used for price prediction in six districts of Sri Lanka region [17]. The system produced 89.66% of accuracy for a given data set. The strategy gives the idea of working season wise data to have more accurate crop prediction for each season separately. The analysis and models of predictions are enhanced continuously. One of the enhancements is applying the deep learning algorithms for crop prediction. The extensively used deep learning algorithms for crop prediction are LSTM, CNN and ANN [18]. The document gives an idea that NN-based deep learning algorithms give the best result.

3 Methodology

The season wise crop yield prediction is implemented using linear regression (LR), Bayesian linear regression (BLR), boosted decision tree regression (BDTR), decision forest regression (DFR) and Poisson regression (PR) algorithms. The implementation flow has been represented in Fig. 1. The flow has been implementation separately for crop prediction and rainfall prediction.

3.1 Linear Regression

The simplest linear regression algorithm assumes the relationship between input and output variables. The input and output variables can be used vice versa due to its correlation coefficient between them. The output (Y) prediction from the input (X) is shown in Eq. 1.

$$Y = m + nX \tag{1}$$

The initial iteration starts with a random m and n values. These values are updated to minimize the cost function of Eq. 2 which minimize the root mean square error (RMSE) and fix the best linear line.

$$\text{Cost Function}(J) = 1/n \sum_{k=1}^{n} \sum_{k=1}^{n} (p_k - y_k)^2 \tag{2}$$

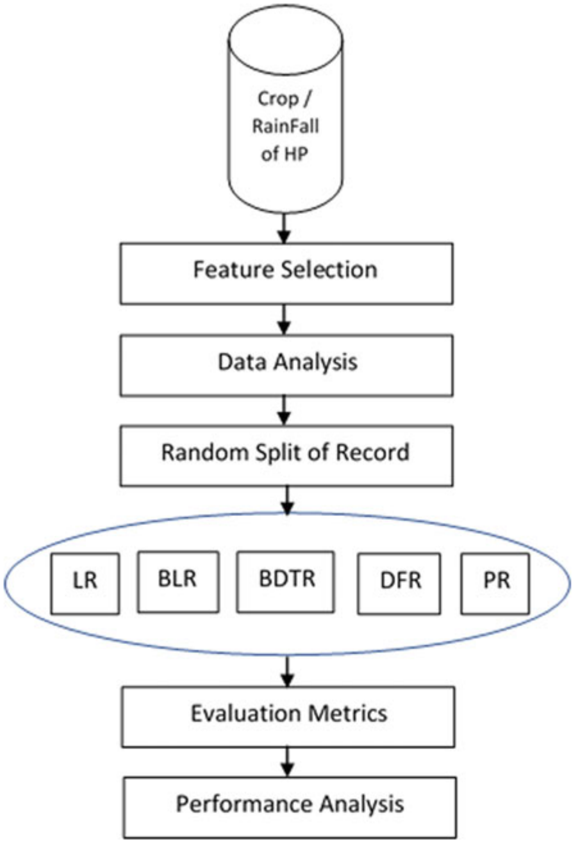


Fig. 1 Implementation flow

3.2 Bayesian Linear Regression

The probability distribution is rather than point distribution. The output value y is generated from the Gaussian distribution characterized by mean and variant. BLR finds the posterior distribution rather than finding the single best value. Posterior distribution in Eq. 3 of the model is dependent on training input and output.

$$p(\beta|y, x) = \frac{p(y|B, x) * P(\beta|x)}{p(y|x)} \tag{3}$$

3.3 Boosted Decision Tree Regression

The BDTR is an enhanced model of DTR, it built a series of trees. The boosted tree depends on the previous tree. The predefined error function corrects the error and improves the tree to increase the accuracy and reduce the error in the next level of tree. Number of branches and the number of trees are directly proportional to the accuracy of prediction. But more trees increase the running time of the model. The loss function is the distance between prediction and the actual value that is identified using Eq. 4. The minimum objective function in Eq. 5 is identified using loss function and regulation function.

$$f(l) = \mathrm{d}\big(\hat{y}_i, y_i\big) \tag{4}$$

$$0(x) = \Sigma_{\dot{L}} f(l) + \Sigma_t f(t) \tag{5}$$

3.4 Decision Forest Regression

The model is an ensemble of decision trees, in which prediction of each tree is given as Gaussian distribution. This category of regression is called as nonparametric models, and it executes as order of simple tests for all cases, navigating up to final result. It can achieve joined features in selection and classification.

3.5 Poisson Regression

Log linear model finds the maximum likelihood to identify the independent variable to predict and observe the dependent target variable. The maximum likelihood found using Eq. 6 is probability-based mass function.

$$\theta = \frac{\mathrm{e}^{y^{\theta' x}} \mathrm{e}^{-\mathrm{e}\theta^1_x}}{y!} \tag{6}$$

The likelihood of multivariate behaviour could be attained by finding the exponential product of each independent variable. The proposed methodology finds the multivariate likelihood using Eq. 7.

$$p(y|x_0 \ldots xm; \theta) = \prod_{i=1}^{m} = \frac{\mathrm{e}^{y} \mathrm{i}^{\theta^1 x} i_e - \mathrm{e}\theta^1 x_i}{y_i!} \tag{7}$$

4 Evaluation Metrics

The performance of the each model has been analysed using standard evaluation metrics which are applicable for prediction.

Mean absolute error: The median value of the complete value of the error for the regression model's predictions. The value range starts from 0 and ends with infinity value. Smaller value indicates the better performance model.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{\Pi} |y_i - x_i|}{n} \tag{8}$$

Root Mean Square Error: RMSE is calculated as the square root of the mean of the squared differences between real outcomes and forecast values.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - x_i)^2}{n}} \tag{9}$$

Mean squared error: MSE is all the error values, squares and its average. The value range starts from 0 and ends with infinity value. Smaller value indicates the better performance model.

$$\mathrm{MSE} = \frac{\sum_{i=1}^{n} (y_i - x_i)^2}{n} \tag{10}$$

Coefficient of Determination (R2): Represents the percentage of the variance in the outcome variable that the model and its predictor variables. The value range starts from 0 and ends with 1, higher value indicates the better performance model.

Relative Squared Error (RSE): It evaluates the performance with respect to the simple forecaster performance. The RSE standardizes the total squared error of the verified model and divides it by the total squared error of the simple predictor. The values near to zero is good.

$$\mathrm{RSE} = \sum_{i=1}^{n} \left(y_i - y_i'\right)^2 \Big/ \sum_{i=1}^{n} \left(z_i - z_i'\right)^2 \tag{11}$$

Relative Absolute Error (RAE): It is the same as RSE, and it uses absolute error instead of squared error. The value range starts from 0 and ends with infinity value. Smaller value indicates the better performance model.

$$\mathrm{RAE} = \sum_{i=1}^{n} \left|y_i - y_i'\right| \Big/ \sum_{i=1}^{n} \left|z_i - z_i'\right| \tag{12}$$

5 Data Analysis

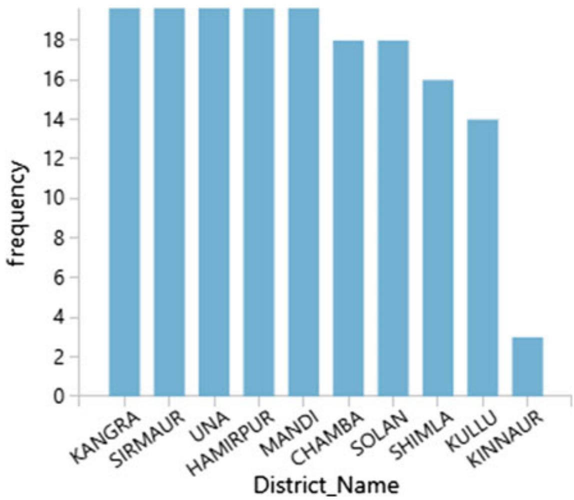
The rainfall and season wise crop production data for 5 years from www.data.gov.in have been taken for the study. Two crop seasons Rabi and Kharif of Himachal Pradesh from 1998 to 2002 are taken. Rabi is the winter season that falls between October and March. Kharif is the monsoon season that falls between July and October. In each season, there are 1826 records available among which 80% records are taken to train the model and 20% are taken to test the model. The data set consists of different district crop production statistics of Himachal Pradesh. Different selections of crops are cultivated from various states of Himachal Pradesh.

The graph in Fig. 2 shows the crop status of area wise different districts in Himachal Pradesh. Kangra, Sirmaur, Una, Hamipur and Mandi are the top three crop levels.

The graph shows the different crops of different districts in area wise cultivation in Himachal Pradesh. The dry chilies is the highest crop produced in Himachal Pradesh, and Garlic is the second highest crop; sugarcane potatoes are next to that in the state. Cotton and tobacco are the lowest crops produced by the state. The same has been visualized in Fig. 3.

The graph in Fig. 4 shows the production versus average rainfall of the state. The known fact is rainfall increases automatically, production also increases and both are directly proportional to each other.

Fig. 2 District wise crop status



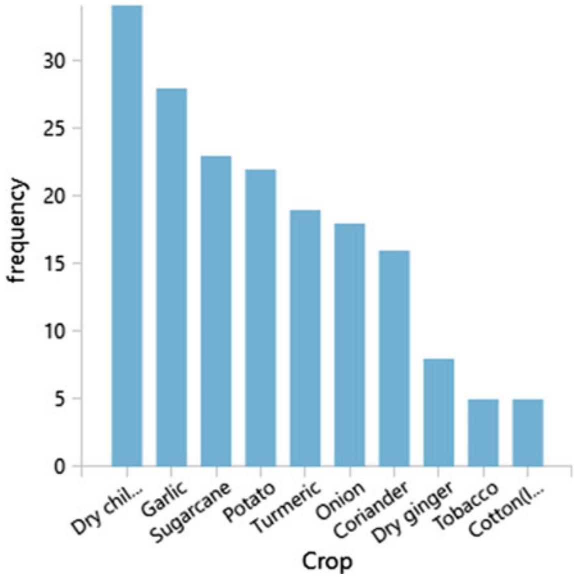


Fig. 3 Different crops versus cultivation area

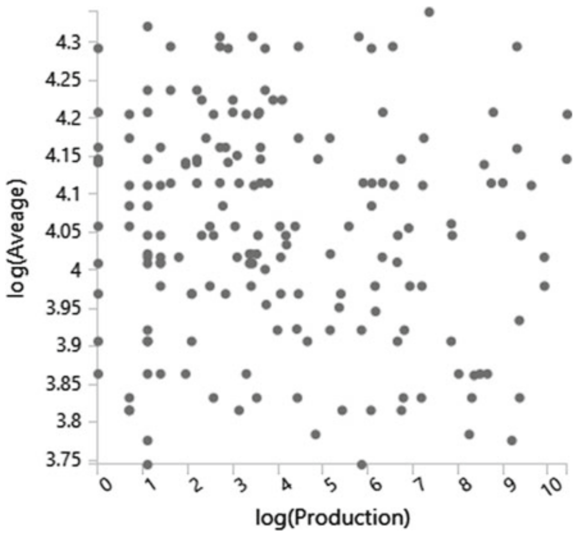


Fig. 4 Production versus rainfall

6 Season Wise Comparative Models Results and Discussion

The two major crop seasons Kharif and Rabi of Himachal Pradesh are considered for prediction models. Agricultural production is mainly dependent on rainfall, so the prediction is carried out using rainfall and crop production.

In Rabi, crop predictions results show that the boosted decision tree regression model is performing well in terms of RAE and RSE. The coefficient determinant ranges between 0.7 and 1. The 80% of Coe proves that it is a good model. Linear regression model and Bayesian linear regression are performing the same. Compared to all other models, the decision tree forest model is the worst performer, the same shown in Fig. 5.

The experimental results of rainfall predictions of Rabi show that the linear regression and Bayesian linear regression models are performing well with 96% of accuracy. Following to that boosted decision tree regression model is carrying out well with 95% accuracy. The same has been visualized in Fig. 6. The Poisson model is not performing well in this prediction model when compare to all the models, the Bayesian linear regression and linear regression models are performing well.

The crop predictions of Kharif experimental results of Bayesian linear regression models are carrying out well with 91.72% of accuracy. Following to that boosted decision tree regression model is performing well with 91.69% accuracy. The Poisson model is not performing well in this prediction model. Compared to all the models, the Bayesian linear regression and linear regression model are performing well. It is visualized in Fig. 7.

The rainfall predictions of Kharif experimental results shows that linear regression and Bayesian linear regression models are performing well with 96% of accuracy. Next to that, boosted decision tree regression and decision forest regression are performing well with 95% accuracy. The Poisson model is not performing well in this prediction model when compare to all the models, the Bayesian linear regression and linear regression models are performing well Fig. 8.

The experimental results of the Rabi and Kharif seasons crop prediction show that the boosted decision tree regression model is performing well compared to all

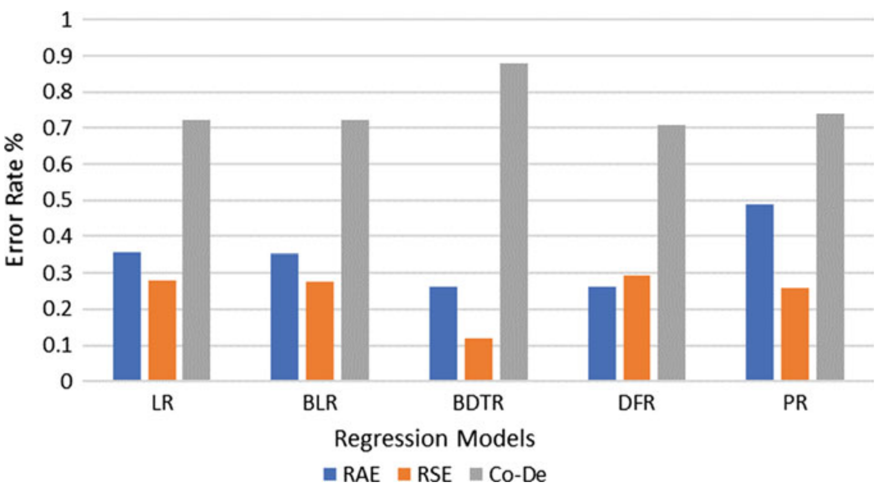


Fig. 5 Crop predictions for Rabi season

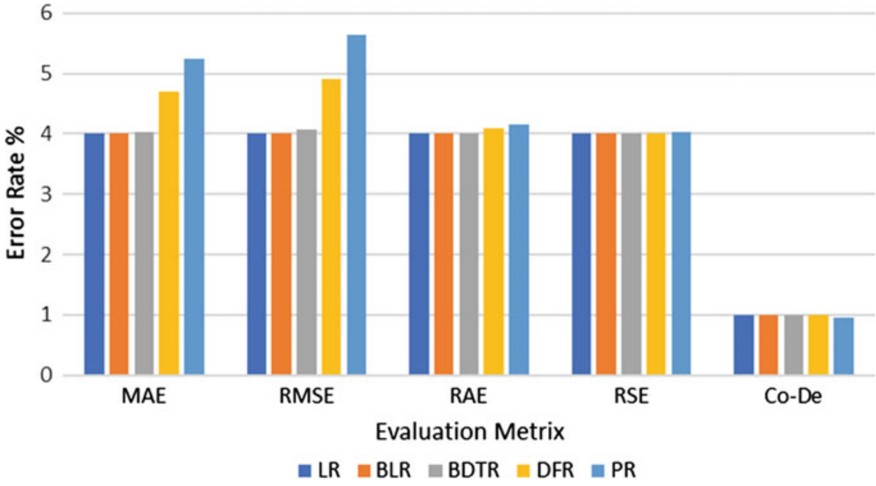


Fig. 6 Rainfall predictions for Rabi season

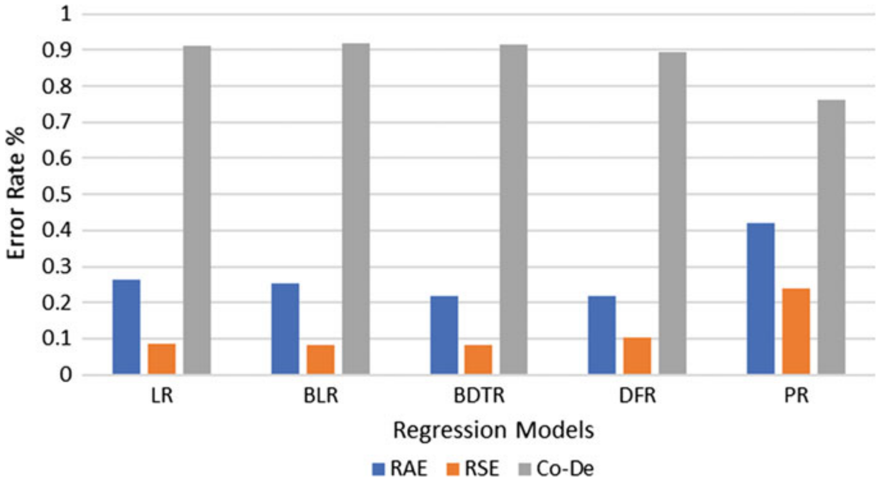


Fig. 7 Crop predictions for Kharif season

other referred models. The rainfall predictions of Rabi and Kharif seasons show that the linear regression and Bayesian linear regression models are performing well compared to all other referred models.

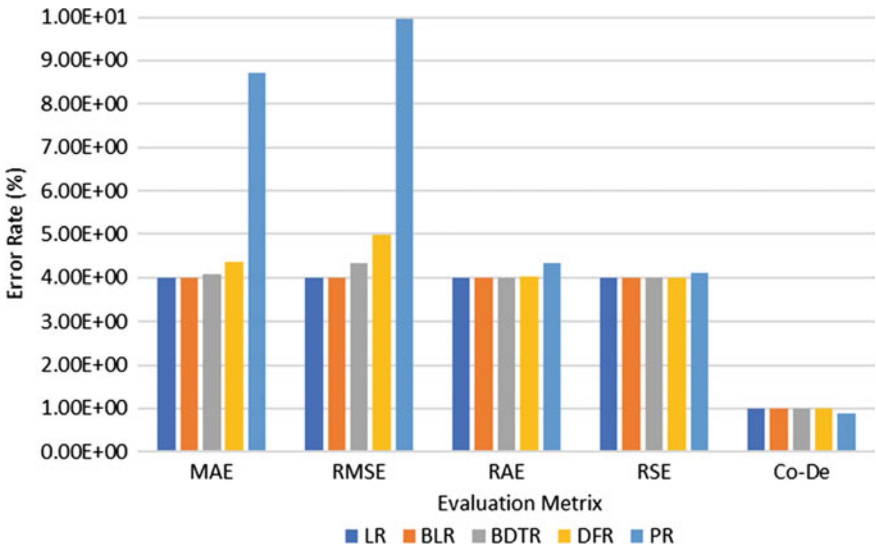


Fig. 8 Rainfall predictions for Kharif season

7 Conclusion

The agricultural prediction models are relying on the data. The changing nature of climate is a challenging parameter in the prediction model. The ML techniques are one of the supporting procedures that support the model in, predicting the best result. The boosted decision tree regression, linear regression, decision forest regression, Bayesian linear regression and Poisson regression are applied on the data to predict the crop production of various crops in Himachal Pradesh region. The study has been done separately on each crop of winter crops and monsoon crops. The investigational outcomes demonstrate that the boosted decision tree regression model is generating improved results in terms of RMSE, RAE, RSE and coefficient of determination compared to all other methods for crop prediction. The results prove that the linear regression and Bayesian linear regression model are producing better results compared to all other methods for rainfall prediction.

References

1. Dharani MK, Thamilselvan R, Natesan P, Kalaivaani PCD, Santhoshkumar S (2021) Review on crop prediction using deep learning techniques. J Phys Conf Ser 1767:1–10
2. Pandith V, Kour H, Singh S, Manhas J (2020) And Vinod Sharma, performance evaluation of machine learning techniques for mustard crop yield prediction from soil analysis. J Sci Res 64(2):394–398

3. Balakrishnan N, Muthukumarasamy G (2016) Crop production-ensemble machine learning model for prediction. Int J Comp Sci Softw Eng (IJCSSE) 5(7):148–153
4. Naveen Kumar PR, Manikanta KB, Venkatesh BY, Naveen Kumar R, Patil AM (2020) Smart agricultural crop prediction using machine learning. J Xi'an Univ Architect Technol 7(5):1606–1614
5. Gabhane Srushti N, Shaikh Naushinnaaz H, Sadavarte Shivani M, Khan Huda A, Waghmare AI (2020) Crop yield prediction to maximize profit using machine learning. Int Res J Modern Eng Technol Sci 02(06):176–182
6. Nandhini N, Gowri Shankar J (2020) Prediction of crop growth using machine learning based on seed features. Ictact J Soft Comput 11(01):2232–2236
7. Kalimuthu M, Vaishnavi P, Kishore M (2020) Crop prediction using machine learning. In: Proceedings of the third international conference on smart systems and inventive technology (ICSSIT 2020) IEEE Xplore Part Number: Cfp20p17-Art, pp 926–932. ISBN: 978-1-7281-5821-1
8. Thomas KT, Varsha S, Saji MM, Varghese L, Thomas EJ (2020) Crop prediction using machine learning. Int J Future Gen Commun Netw 13(3):1896–1901
9. Nischitha K, Vishwakarma D, Ashwini, Mahendra N, Manjuraju MR (2020) Crop prediction using machine learning approaches. Int J Eng Res Technol (IJERT) 9(08):23–26
10. Zingade DS, Buchade O, Mehta N, Ghodekar S, Mehta C (2017) Crop prediction system using machine learning. Int J Adv Eng Res Develop Spec Issue Rec Trends Data Eng 4(Special Issue 5):1–6
11. Kumar A, Kumar N, Vats V (2018) Efficient crop yield prediction using machine learning algorithms. Int Res J Eng Technol 05(06):3151–3159
12. Patil P, Panpatil V, Kokate S (2020) Crop prediction system using machine learning algorithms. Int Res J Eng Technol (IRJET) 07(02):748–753
13. Manjula Josephine B, Ruth Ramya K, Rama Rao KVSN, Kuchibhotla S, Venkata Bala Kishore P, Rahamathulla S (2020) Crop yield prediction using machine learning. Int J Sci Technol Res 9(02):2102–2106
14. Palanivel K, Surianarayanan C (2019) An approach for prediction of crop yield using machine learning and big data techniques. Int J Comp Eng Technol (IJCET) 10(03):110–118
15. Jiang Z, Liu C, Hendricks NP, Subramanian BG, Hayes DJ, Sarkar S (2018) Predicting county level corn yields using deep long short term memory models. arXiv preprint arXiv:1805.12044
16. Alhnaity B, Pearson S, Leontidis G, Kollias S (2019) Using deep learning to predict plant growth and yield in greenhouse environments. Int Symp Adv Technol Manage Innov Greenhouses: GreenSys2019 1296:425–432
17. Selvanayagam T, Suganya S, Palendrarajah P, Manogarathash MP, Gamage A, Kasthurirathna D (2019) Agro-genius: crop prediction using machine learning, Vol 4, Issue 10, Oct 2019, pp 243–249
18. Van Klompenburg T, Kassahun A, Catal C, Learning CYPUM (2020) A systematic literature review. Comp Electron Agric 177:1–18

Fog-Enabled Energy-Efficient Smart Street Lamps using Machine Learning and IoT

J. Angela Jennifa Sujana, R. Vennita Raj, and V. K. Raja Priya

Abstract Street lighting becomes an imperative component for road safety and crime prevention. The main factor in street lighting is energy consumption. Light emitting diode (LED) lighting technology assures energy savings and even dispersion of light intensity. The modularization of LEDs considers street light maintenance and replacement. The public street lamps consume more energy without the lack of energy-efficient management. The proposed system implements energy-efficient management over fog servers by automating the adjustment of luminance in street lights. The alteration of brightness over the existence of pedestrians and vehicles using IoT sensors and machine learning magnifies the efficient power consumption at higher levels. The environment provides the system to recognize the patterns by employing the gradient boosting algorithm. Based on the obtained knowledge, the luminance of street lamps is adjusted. The fog-enabled IoT framework contributes to various performances in quality, accuracy, power consumption, and reduced latency.

Keywords Fog computing · Internet of things · Machine learning

1 Introduction

Nowadays, the Internet of things (IoT) is gaining popularity over various the industries due to its automation and energy consumption. The process of automation is still enhanced with intelligence over the system to achieve an energy-efficient model. The street lighting in cities increases security over pedestrians, riders, and drivers in

J. Angela Jennifa Sujana (✉)
Department of Artificial Intelligence and Data Science, Mepco Schlenk Engineering College, Sivakasi, India
e-mail: ang_jenefa@mepcoeng.ac.in

R. Vennita Raj · V. K. Raja Priya
Department of IT, Mepco Schlenk Engineering College, Sivakasi, India
e-mail: rajapriya@mepcoeng.ac.in

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_21

[1, 2]. The major factors involved in efficient street light are less power consumption in light intensity [3] and efficient communication network management between sensors [4]. However, the street lamps face some challenges,

1. High energy consumption—Street lamps are essential in promoting security over urban/rural areas. But street lamps consume large amounts of energy, where in the future, human resources may face increasing demand. Thus, reducing or efficient usage of street lights plays a vital role in the construction of well-organized smart cities.
2. Lack of management—The management of street lights over theft and robbery is still a big deal. The absence or damage of street lights only comes to knowledge when visiting the actual sight, which results in poor lighting over the system.
3. Maintenance—The maintenance over a large network of street lamps is tedious. Normally, manual maintenance took at regular intervals of 12 months. But the prior information in draining resources helps to improve effective street lighting.
4. Environmental impact—The lower the usage of non-renewable resources will lower the impact on the environment.

The proposed system designed 'Smart Street Lamp' using IoT and machine learning to overcome the challenges and efficient management by fog computing. The smartness is developed using sensors where the light glow comes into action based on demand. The real-time detection of movement makes the necessary lighting under fog computing [5]. The dim glow of street lamps on the lack of pedestrians or vehicles conserves huge amounts of energy, which thrives on energy-efficient management. The intelligence over the system adds to the autonomous maintenance over the system. The objectives in 'Smart Street Lamp' design are:

- To design energy conserving units in streetlamps by adjusting the lightness based on requirement.
- To improve the efficient management in streetlights by tracking the states of streetlamps over fog servers.
- To integrate intelligence with smartness using machine learning and IoT.
- To protect the environment with reduced wastage of energy by building energy-efficient 'Smart Street Lamps'.

2 Related Work

The decentralized infrastructure of fog computing provides services closer to the user and effective in real-time applications. The fog computing platform endures with reduced latency, network bandwidth, and improved user experience [5]. The two ways of control of street light systems are central control and distributed control. The distributed control provides better communication and management on streetlights. A street light management system (SLMS) is developing to improve dynamic luminosity, fine control, and management and independent warning in the abnormal states. The intelligent traffic flow tracks the traffic congestion based on Q-learning in

[7]. The self-adaptability and independence in learning help in the control of traffic lights under fog computing. The feasibility of fog computing over complex, heterogeneous, and dynamic structures lacks in the management of relevant processes. The automation and intelligence over the process are increased with the application of machine learning to fog computing in [8]. The prediction or decision making based on supervised learning further enhances the process of machine learning with less human intervention. The knowledge extraction from learning enhances the management of the resource, the security, and accuracy in fog computing. Moreover, the various issues, concepts, applications, and challenges on the implementation of fog computing and machine learning are reviewed.

The distributed network protocol is analyzed by machine learning in [9]. The learning is based on the prediction of network quality parameters for a fog computing service. The realistic fog-based sandbox is deployed with the image processing ensemble of services. The objective of implementing learning in fog infrastructure is to improve the adaptation of the model more efficiently to latency-sensitive environments. The process of machine learning techniques such as k-means clustering is developed to process detecting diabetes patients on a fog computing framework named FogLearn in [10]. FogLearn framework acts as an edge computing environment in processing geospatial and medical data images. The cloud assisted fog computing emergence to various computational paradigms such as mobile cloud computing (MCC) and mobile edge computing (MEC). The fog layer provides computing nearer to the edge devices and effective in the real-time applications such as geospatial application, healthcare, smart city, smart grid, smart home, etc.

3 Smart Street Lamp

The Smart Street Lamp consists of 1) the smart lamp—street lamp equipped with sensors to track the presence of vehicles or pedestrian. 2) Communication infrastructure—the network of communication between each lamp and control center. 3) Control and management center—monitor and control the streetlamps. The 'Smart Street Lamp' communicates with one another and also with the management center (see Fig. 1).

3.1 The Smart Lamp

The 'Smart Street Lamp' is furnished with sensors such as the photoresistor, motion detection, and the range finder to track the traffic flow. The solar powered street lamps encourage the use of renewable energy resources. The light dependent resistor (LDR) sensor detects the decline of solar radiation which sets street lamps into a dim glow state. The required modification in lightness level comes as a result of measured movement by sensors. The passive infrared (PIR)-based motion detectors

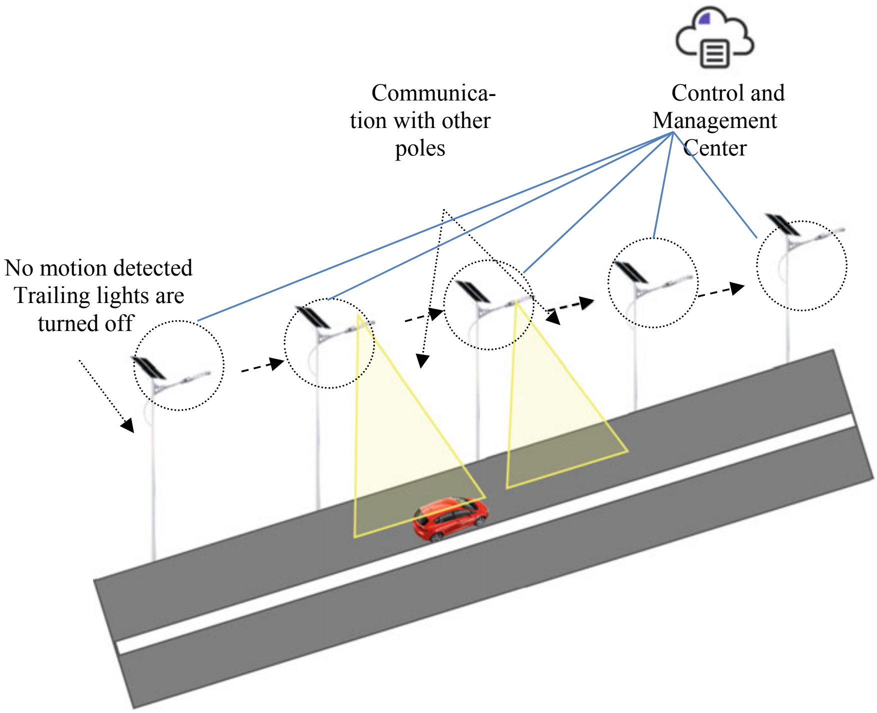


Fig. 1 Example of 'Smart Street Lamp' scenario

measure infrared light refracted from the objects, whereas rangefinder measures the sound waves reflected and calculates the distance of the object location. The interactive results from sensors aid in alteration of essential lightness level. The energy conserving state is attained with the dim glow of street lamps in the absence of street activity.

3.2 Communication Infrastructure

Smart lamps communicate with succeeding poles on indication of movement in traffic activity. Based on the gained knowledge from trailing lights, the successor modifies its lightness. There are two levels of communication: 1) edge level communication and 2) fog level communication.

The edge level communication defines the communication between the edge nodes. When the sensing unit in the streetlamp predicts the traffic activity it indicates the motion detection to the succeeding lamp. The announcement allows the successor to alter its lightness even before its sensing unit predicts the movement. The process mainly improves the smartness among the networks. The fog level communication states the communication from edge devices to the fog server. The collection of interactive results from the smart lamps is sent to the fog server. The fog server controls and manages the streetlamps based on the intelligence of machine learning.

3.3 *Control and Management Center*

The control and management center consists of fog nodes and fog data services. The fog nodes are a special-purpose computing system which handles the collection of large volumes of data. The computing system forwards the data to obtain services of machine learning analysis from the fog data services.

Machine Learning. The fog data services provides services over the requisition of job/task. The proposed fog architecture with 'Smart Street Lamp' consists of modules of services provided by fog data services (see Fig. 2):

- Data manager
- Data visualization
- Cloud integrator.

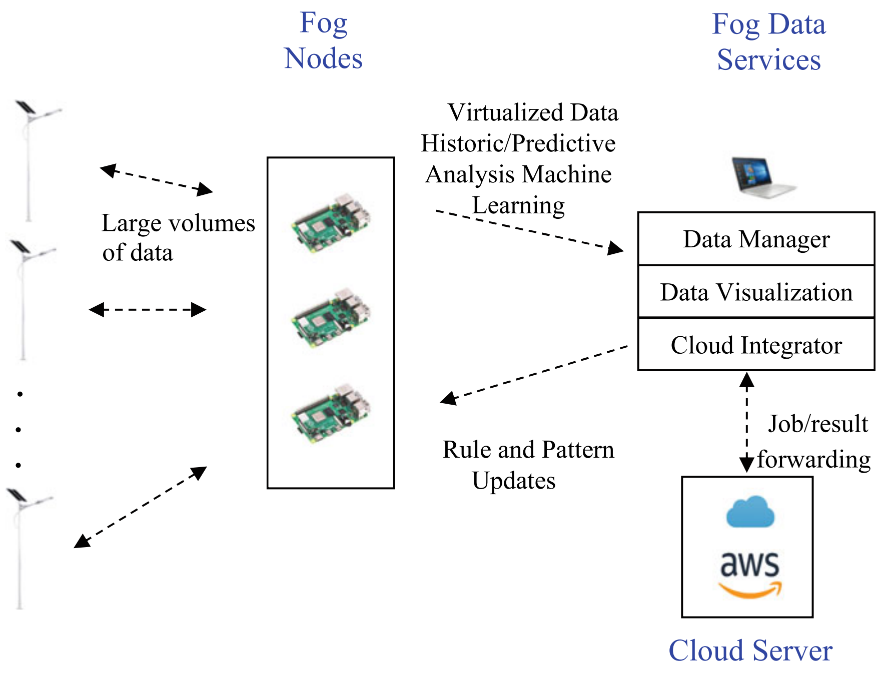


Fig. 2 Proposed fog architecture with 'Smart Street Lamp'

The data manager module receives the request and implements data mining analytics such as preprocessing on raw data. The flexible management module also provides predictive analysis on incoming historic data using machine learning. Machine learning automates the analytical model by analyzing data and making decisions without human demand. The learning undergoes gradient boosting machine learning technique. Gradient boosting model is generated by combining various simple models into a single composite model. The simple models state the generation of various decision trees as weak learners. The steps involved in generating gradient boosting technique are as follows:

i. Select a weak learner—The constraints of weak learners can be with the number of nodes, leaf nodes and layers.
ii. Generate loss function—Loss function based on gradient descent approach is used as an optimization technique in generating better models.
iii. Use an additive model—The generated decision trees are added one at a time with the existing trees in the model.
iv. Minimize the loss function—The loss function is used to minimize the error/loss while adding decision trees.

The service by the fog server includes the insights of data with several trends and patterns by the module data visualization. The cloud integrator handles the interactivity between the fog server and the cloud server.

4 Design and Implementation

4.1 The Smart Lamp Construction

The smart lamp defines the streetlamps equipped with sensors such as light dependent resistor (LDR) sensor, ultrasonic sensor, and motion detection sensor to detect the traffic activity. The lightness in a streetlamp is adjusted based on its predicting values. Below representation describes the pseudocode for implementing the smart lamp. The proposed 'Smart Street Lamp' performs dynamic behavior with adjusting brightness (see Fig. 3).

PSEUDOCODE

```
//Define the level of brightness by pulse width modulation (PWM)
    LEVEL1 63
    LEVEL2 127
    LEVEL3 191
    LEVEL4 255
    //Define the total power supply
    VOLT 5
    Define the receiver (streetLamp2) MAC address
```

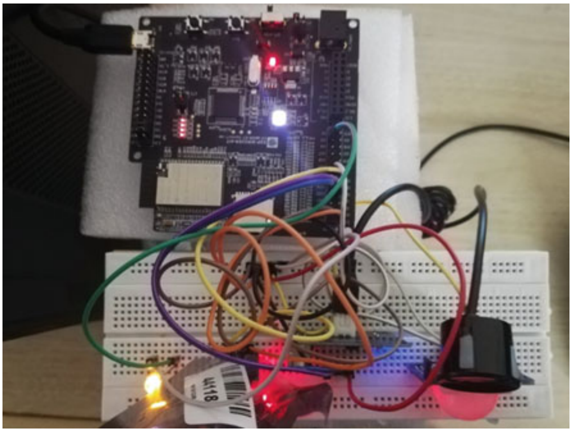


Fig. 3 Experimental setup

Repeat forever:

Read the intensity of light value
Read the duration of signal from objects
Calculate the distance of the object
if intensity of light > = 300

then turn on LED with LEVEL1 brightness

else if intensity of light > = 300 AND motion detection = = 1

turn on LED with LEVEL2 brightness.
Send motion detection value to streetLamp2.

end if
else if intensity of light > = 300 AND distance < = 70 cm

turn on LED with LEVEL3 brightness.
Send distance value to streetLamp2.

end if
else if intensity of light > = 300 AND motion detection = = 1AND distance < = 70 cm

turn on LED with LEVEL4 brightness
Send motion detection value AND distance value to streetLamp2

end if
else turn off LED
power consumed = (level of brightness/256) * VOLT

4.2 Predictive Analysis

A decision tree is constructed by partitioning the feature space into disjoint regions R_j which can be formally expressed as

$$T(x, \gamma) = \sum_{j=1}^{J} \beta_j I\left(x \in R_j\right)$$

where $I\left(x \in R_j\right)$ indicates as a value of 1 if $x \in R_j$ or 0 otherwise.

β_j is the optimal constant for each region.

γ is used to find the region R_j.

The additive model is a sum of M trees

$$f_M(x) = \sum_{m=1}^{M} T(x, \gamma_m)$$

The gradient descent-based loss function can be expressed as

$$r = -\left[\frac{\partial L(y, f(x))}{\partial f(x)}\right]$$

The updated model function with minimized loss function can be calculated as

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I\left(x \in R_{jm}\right)$$

PSEUDOCODE
#Input
Input data $(x, y)_{i=1}^{N}$
Number of iterations M
Loss function L(y, f)
Base learner model $T(x, \gamma)$
#Function
Initialize the function $f_0(x)$ as min of $L(y, \gamma)$
for m = 1 to M

for i = 1 to N

Compute gradient descent-based loss function

end for
Find terminal regions R_{jm} by fitting a regression tree to the loss function
Compute optimal constants for each region

Table 1 Sample of Smart Street Lamp dataset

S. No	LDR	Ultrasonic	PIR	Level of brightness
0	316	46	1	5
1	310	70	1	5
2	138	60	0	1
3	146	61	1	1
4	275	40	1	1
5	144	788	0	1

Update the model function

end for
Calculate boosted tree model as a sum of M trees

5 Performance Evaluation

The proposed system is implemented and deployed in a fog computing environment to develop a model in the prediction of altered brightness in streetlamps. The automation and knowledge-based learning is enhanced with gradient boosting machine learning technique. Table 1 defines the first five rows of the 'Smart Street Lamp' dataset.

The performance of the machine learning algorithm is estimated by the k-fold cross-validation procedure on multiple 'folds' of the dataset. The general procedure of k-fold cross-validation is first split the dataset into k groups. Then, for each group consider that group set data as test data and remaining groups as training data set. Then, the model is fit on the training set and evaluate the model on the test set. Then, the evaluation score is retained by discarding the model. Then, the skill of the model is summarized with the model evaluation scores.

The evaluation of gradient boosting classifier reports the mean classification accuracy of 0.975 on new data using tenfold cross-validation (see Fig. 4). The distribution of scores is summarized by a box and whisker plot (see Fig. 5). Each plot represents the median of the distribution as orange line and arithmetic mean as green triangle. The box and whisker plot representation helps in choosing the number of trees set.

6 Conclusion

The proposed 'Smart Street Lamp' model automates the lightness in streetlights based on machine learning. The automation enhances the process of self-decision making over the learned knowledge from real-time data. The fog-based machine

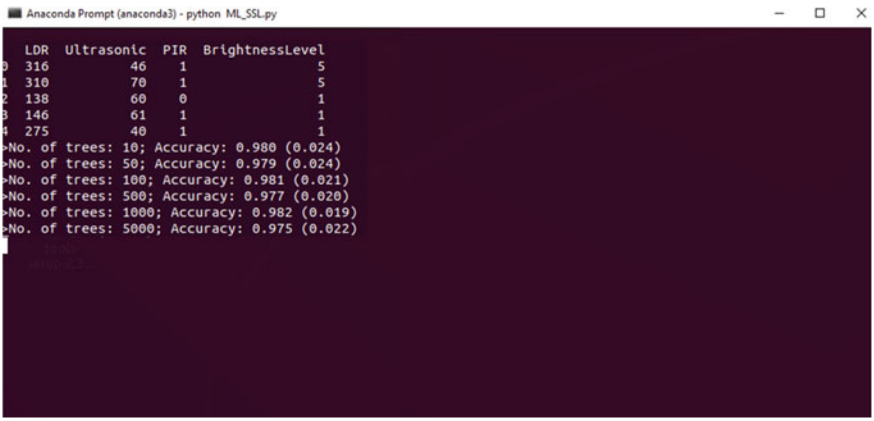


Fig. 4 Model accuracy on different set of trees

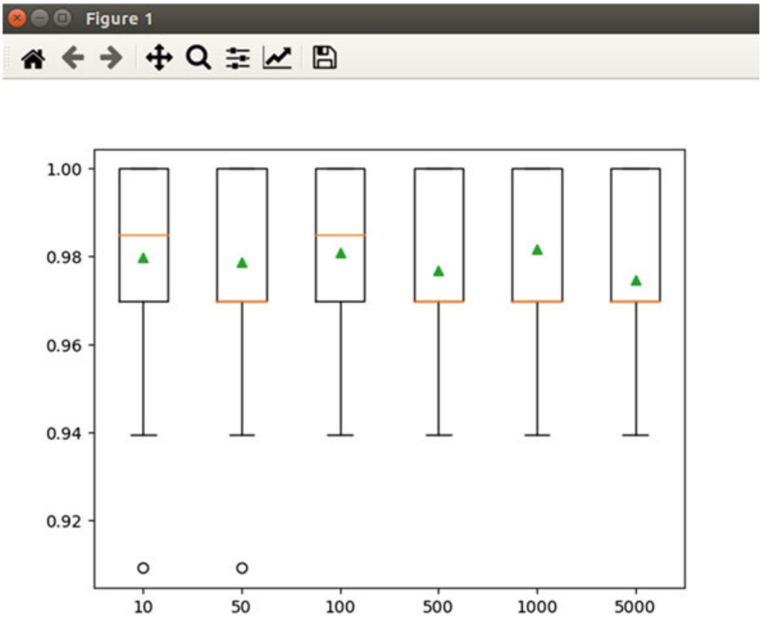


Fig. 5 Box and whisker plot representation on the distribution of scores

learning models reduce latency, power consumption, and network bandwidth. The proposed system efficiently manages the control and maintenance over the smart street light system which reduces periodic inspection. In the future, the proposed system can be further extended on various applications such as agriculture, in smart cities—traffic management, environmental monitoring, parking, and so on.

Acknowledgements The authors acknowledge the support and grant for this work under the project titled 'Energy Aware Smart Street Lamp Design for Autonomous Maintenance using AI and Fog Computing' (vide No.: SP/YO/2019/1304(G), dated 20.05.2020) from Department of Science & Technology, SEED Division, Ministry of Science & Technology.

References

1. L. P. Maguluri, Yaswanth S. V. Sorapalli, Lokesh kumar Nakkala, Venkat Tallari: Smart Street Lights using IoT. 3rd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT) 126–131 (2018).
2. Siddharthan Chitra Suseedran, Nanda Kishore B, Josephus Andrew, M. S. Bennet Prabha: Smart Street lighting System. 3rd International Conference on Communication and Electronics Systems (ICCES) 630–633 (2018).
3. Akansha kaul, Tina chaudhary, Nathi Ram Chauhan: Design and Control of Street Light Intensity for Fog and Rain. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering 6(7), 5580–5587 (2017).
4. Ngo Thanh Tung, Le Minh Phuong, Nguyen Minh Huy: Development and Implementation of Smart Street Lighting System based on Lora Technology. International Symposium on Electrical and Electronics Engineering (ISEE), 328–333 (2019).
5. Noriaki Yoshiura, Yusaku Fujii, Naoya Ohta: 13th International Symposium on Communications and Information Technologies (ISCIT) 633–637 (2013).
6. Rekha T (2020) P Sreenivasulu: Fogging Based Street Light Management System for Smart City. Journal of Xi'an University of Architecture & Technology 12(11):494–502
7. Qin H, Zhang H (2021) Intelligent traffic light under fog computing platform in data control of real-time traffic flow. J Supercomput 77:4461–4483
8. Karrar Hameed Abdulkareem (2019) Mazin Abed Mohammed, Saraswathy Shamini Gunasekaran, Mohammed Naseer Al-Mhiqani, Ammar Awad Mutlag, Salama A. Mostafa, Nabeel Salih Ali, Dheyaa Ahmed Ibrahim: A Review of Fog Computing and Machine Learning: IEEE Access 7:153123–153140
9. Jon Patman; Meshal Alfarhood; Soliman Islam; Mauro Lemus; Prasad Calyam; Kannappan Palaniappan: IEEE INFOCOM 2018 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS) 790–795 (2018).
10. Robindra K. Barik, Rojalina Priyadarshini, Harishchandra Dubey, Vinay Kumar, and Kunal Mankodiya: FogLearn: Leveraging Fog-based Machine Learning for Smart System Big Data Analytics. International Conference on Energy. International Journal of Fog Computing 1(1) 15–34 (2018).

Sensing of Nitrogen and Temperature Using Chlorophyll Maps in Precision Agriculture

Sahadevan Sakthipriya and Ramu Naresh

Abstract Precision agriculture is a technique of agricultural management that makes use of spatial data. It is ensuring nutrient status of crop which is critical for monitoring plant production, vegetation health, soil health and management nutrient optimisation initiatives. Using multi-spectral Landsat as a target for crop strength sensing in precision agriculture, this study maps the spatial distribution of chlorophyll level and yield maps for fields under corn (Zea mays) and wheat (Triticum aestivum) maps to different quantities of nitrogen and temperature fertiliser application. An analytical of spectral devices (SD 500–2600 nm) was used to sampling hyper-spectral reflectance readings. Using Landsat-9 (30m) dependence data at the subfield size and integrated canopy and leaf radioactive transport models. The chlorophyll gauge was monitoring to stress level for management. This research study shows PA model contained by leaf chlorophyll as a predictor of crop temperature and nitrogen stress levels and multi-spectral dependence satellite data, SPAD-503 and multi-spectral systems.

Keywords Leaf · Chlorophyll content · Wheat · Corn · Landsat · Nitrogen · Temperature and SPAD

1 Introduction

Precision agriculture, also known as remote sensing agriculture or site-specific crop management, is a type of crop management that focuses on a single location. Crop and soil conditions, crop management, air movement, manpower prices, equipment utilisation, hyper-local weather forecasts and other data are all accessed in real time with this strategy. This study examined the notion of precision agriculture through

S. Sakthipriya
Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Chengalpattu, Chennai, Tamilnadu 603 203, India

R. Naresh (✉)
Department of Networking and Communications, SRM Institute of Science and Technology, Kattankulathur, Chengalpattu, Chennai, Tamilnadu 603 203, India
e-mail: nareshr@srmist.edu.in

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_22

machine learning, as well as the numerous ML algorithms employed and their reliability. Precision agriculture research aims to develop an assessment support system for entire ranch management by detecting, testing and reacting to changes in crops both outside and within the field [1]. The purpose of this study is to investigate several forms of classification and clustering algorithms that are used in predictive agriculture to boost production. It explains how to use several regression approaches to get valuable precision. Harvest output can also be utilised to identify portions of the farm that are unprofitable for a phase of time and can be cumulative for preservation without incurring any economic costs for the manufacturer collision. Reduced applied nitrogen application can help to reduce nitrate and temperature levels, but it is part of a bigger picture that includes urbanisation, mineralisation and rainfall. Over applying nitrogen has a variety of short- and long-term consequences, including economic, environmental, habitat and human consequences [2].

Because of its simplicity and computational efficiency, the undergrowth index (vegetation) is a commonly used and directly forward way of predicting leaf chlorophyll content (LCC) from remote sensing data. However, because index of vegetation-based empirical representation relies on actual facts, they are rarely transferable to last regions, stages of growth or species limiting their use more wide areas. Machine learning (ML) techniques are currently being used as an alternative to LCC estimation because of their ability to produce adaptive and resilient associations. It is worth mentioning that one drawback of machine learning algorithms is that they require training data to make predictions [3]. Improper or insufficient training data selection can cause model variations or even mistakes. Space/without time constraints or the use of local calibrations and radiative transfer models (RTM) yield robust estimations and comparatively accurate of LCC across a variety of canopy structures along with soil environment types. RTMs are frequently used to extract canopy and leaf characteristics using look-up tables (LUTs) [4]. However, leveraging satellite remote sensing data to get LCC from RTMs at broad spatial scales remains an arduous task. During photosynthesis, leaf chlorophyll is essential for catching photons and transferring electrons. In various investigations, we discovered that chlorophyll content levels were a good indicator of leaf photosynthetic capacity. They can also be vital for understanding physiological condition, plant stress, and prospective photosynthesis. As a result, accurate LCC retrieval is vital for assessing plants physiological status and is crucial in the modelling of undergrowth productivities [5, 6].

2 Related Works

2.1 Reflectance Field Measurements of Canopy

An SD "Field Spec Pro spectrometer for Boulder, Analytical of Spectral Devices" (USA) with a range of (360–2600 nm) spectral variety, resolutions of 3 nm (360–1050 nm) spectral and 10 nm (1050–2500 nm) was used to assess reflectance canopy

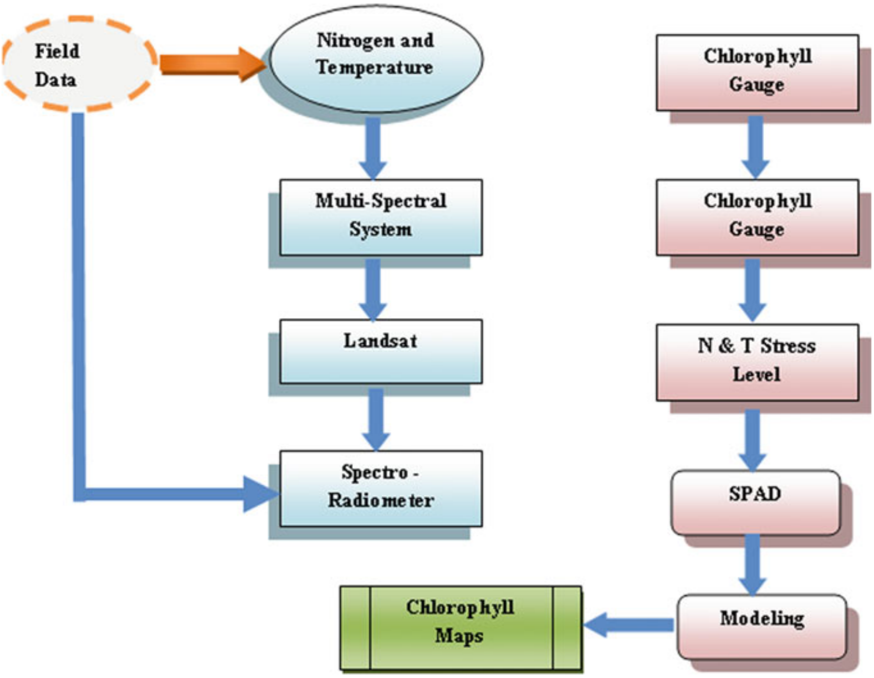


Fig. 1 Proposed work flow diagram

at XTS site. They took the measurements between local time (10:00 and 14:00), at a metres of height 1.3 above the Triticum aestivum (wheat) with a 25 field of vision under clear and cloudless conditions. Each plot's average canopy reflectance has been calculated using 20 individual measurements [7]. Two inter-calibrated Ocean Optics radiometers (USB2000, USA) were used to measure canopy spectral reflectance at the US-Ne (2 and 3) locations, with a spectral range of 401–1100 nm and a spectral motion of 1.5 nm. At an altitude of roughly 5.5 m above the canopy, one radiometer with a 25 field of view measures the crop's upwelling brightness with a hemispherical field of view, the other calculated occasion irradiance downwards. Between 11:00 and 14:00 local time, they performed the measurements under clear skies, and we computed the reflectance using Gitelson's method [8].

Figure 1 shows a work flow of chlorophyll maps for nitrogen and temperature. This shows a figure consisting of different types of data from field data modelling.

2.2 *Chlorophyll Canopy Content*

They collected new wheat folio samples from the peak of the canopy in a 1 m^2 area for each one plot at the XTS site and quickly positioned in an artificial box

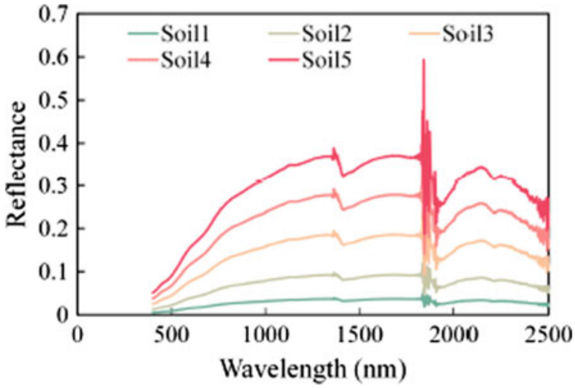


Fig. 2 PROSAIL simulation soil model

containing frost for transportation towards the laboratory. A spectrophotometer was used to determine the chlorophyll concentration. The dry-weight method was used to determine the LAI of the sample leaves [9]. I gathered fresh leaves from corn plants in six small plots (20 m × 20 m) within each site at the US-Ne2 and US-Ne3 sites. LCC was determined using a spectrophotometer after it removed the leaf pigment with 80 per cent acetone. An area metre was used to determine the LAI of sample leaves model Lincoln (LI-3100, USA). We then averaged six plots of LCC and LAI to create site-level values. Increasing the LAI by the LCC yielded the whole chlorophyll limitation for CCC [10].

2.3 PROSAIL-D Model

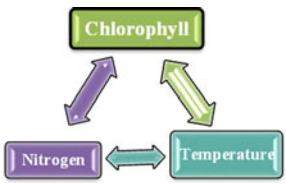
By combining the PROSPECT-D leaf optical characteristics model with the SAIL canopy light intensity model, the PROSAIL-D model was created. To simulate MERIS readings and model the CCC based on VIs, the PROSAIL-D model was employed. It uses seven input parameters to mimic upwards and the downwards hemisphere radiation fluxes between 400 and 2500 nm and outputs the leaf spectral reflectance and transmittance. With a sequence of inputs, the SAIL 5 model parameters are utilised to mimic canopy reflectance. Figure 2 shows PROSAIL-D simulations for input parameters in soil model wavelengths [11, 12].

2.4 Vegetation Indices

To integrate LAI data into the MTCI, several LAI-VIs were chosen. For comparison, the normalised difference vegetation index (NDVI) was utilised, which shows saturation for various plants when the LAI is greater than 2. When calculating the LAI, several studies have changed the NDVI to reduce the effect of saturation. By introducing a linearity change fact LNDVI, they created a linearised NDVI into the NDVI equation. When the vegetation fraction grew too great, a stretched S-NDVI is produced by means of a scaling alteration function to remove saturation. For LAIs of (2.5–5.0), the S-NDVI did not reach saturation when compared to the NDVI. Other spectral indices, besides changed NDVIs, have greater associations with the LAI in heavily vegetated areas. RDVI renormalised difference vegetation indices suggested combining the advance of drawbacks of the difference vegetation index DVI for low LAIs and individuals of the NDVI for elevated LAIs. The RDVI exhibited the strongest positive relationship with the LAI and stayed far from initiating saturation in the incidence of large LAIs, according to researchers who looked at 56 hyper-spectral vegetation indices. MTVI2, a new triangular vegetation index, was developed and found to be the greatest predictor of the LAI. The chlorophyll changes were unaffected by MTVI2 and did not reach dispersion at high LAIs [13].

3 Materials and Methods

It is a remote sensing reserve of exceptional relevance because of its great open access, spatial resolution and long-term repository of information from the Landsat series. However, most physical and experimental methods used for modelling chlorophyll leaf concentration rely on hyper-spectral or narrow band reflectance and utilise the limiting of a broad range of sensors, to install on sensors satellites and SPAD, for example—often model the spectrum at smaller quantity and wider spectral bands [14].



3.1 Study Sites

The grassland sites are in Easthope Township, Ontario, and included corn and wheat two fields. They used 90 per cent of the land in this region for agricultural production, making it a very productive agricultural area. The area's top soils are predominantly silty loams and clay with high fertility nature. Figure 3 shows sample climate growth maps of precision agriculture for wheat and corn. An Analytical of Spectral Devices (SD) Field Spec Pro spectroradiometer (500–2600 nm) was used to

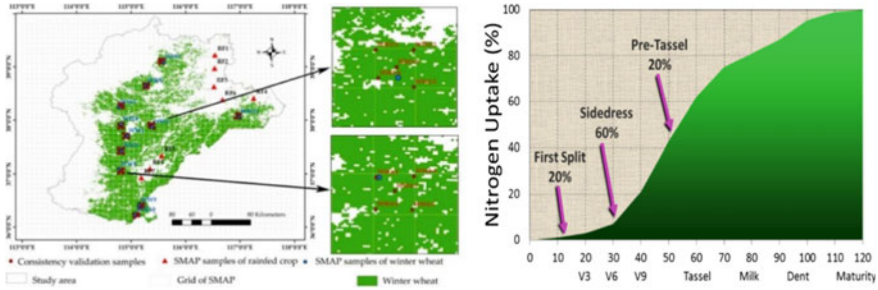


Fig. 3 Sample growth maps of winter wheat and corn

sampling hyper-spectral leaf reflectance readings. Using connected canopy and leaf radioactive transfer models follows a two-step inversion approach estimated content of leaf chlorophyll from Landsat-9 (30 m) dependency data at the subfield scale. The chlorophyll gauge was useful in determining N stress levels for better N management and T stress levels for better T management. It was impossible to isolate the complex of characteristics that contributed to stress as measured by a chlorophyll gauge.

$$P\left({}^{LC}/_{NS,\,TS}\right) = \frac{P\left({}^{NS,\,TS}/_{LC}\right)P(LC)}{P(NS,\,TS)}$$
$$P(LC) \geq P\left({}^{LC}/_{NS}\right)P\left({}^{LC}/_{TS}\right) \tag{1}$$

where

LC—leaf chlorophyll, *NS*—nitrogen stress, *TS*—temperature stress.

3.2 *SPAD-503*

Minolta Corporation, Japan, employed a chlorophyll metre SPAD-503, commonly known as SPAD meter. The chlorophyll metre was retrofitted with an RS233 interface to converse with a wrist band data logger from spectrum technology, with a measure area (2 × 3 mm) and measurement of precision (1.0) SPAD unit at space temperature (room). A backpack global positioning system (AgGPS 122) unit was also linked to the data recorder via another RS232 connector Trimble Corporation. Figure 4 shows *N* of inputs/outputs soil cycling of chlorophyll content.

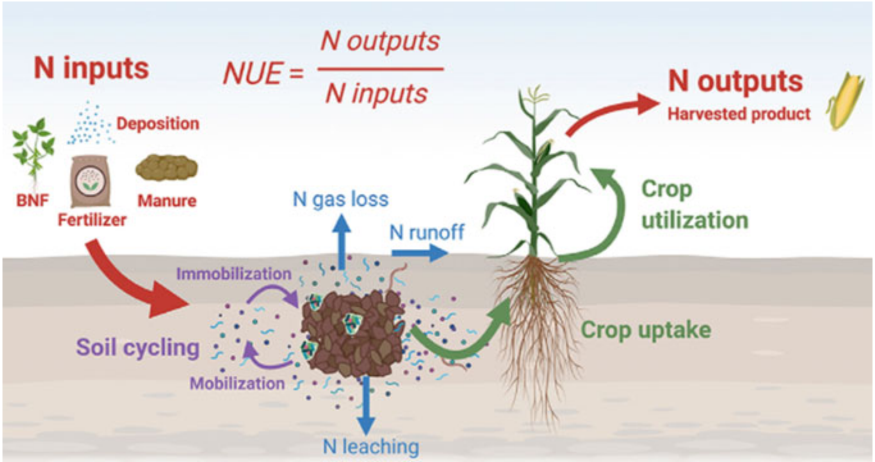


Fig. 4 Soil cycling of SPAD chlorophyll

They used a grouping of GPS and degree of difference GPS correction recipients in the backpack AgGPS 122 system, which had sub-metre accuracy. Radio navigation beacons provided the differential adjustments. It had a single feed line for reliability and a mutual antenna designed for GPS and beacon signals. The display comprised two light-emitting diodes and supplied National Marine Electronics Association (NMEA) output. The following equation is used to calculate the chlorophyll concentration of nitrogen and the temperature meter's transmittance based on green (650 nm) and red (940 nm) wavelengths:

$$S = C \log\left[\frac{TN_t}{LN_o}\right] \tag{2}$$

The levels of chlorophyll are equipment availability, growth stage and trends in all elements that influence when N time is applied. Delaying application N until later than silking is unlikely to result in a yield boost. In irrigated locations, the response to N treatment might begin as early as two days. The amount of nitrogen injected could be between 22 and 42 kg ha^{-1} (20–45 lbs acre). Because dirty soils absorb more nitrogen and temperature ($N\&T$) than other soil textures, the amount of supplied nitrogen would be soil sediment, besides other parameters.

$$N\&T = \left[\frac{\text{Set of Mean value}}{\text{Reading Value}}\right] * 100\% \tag{3}$$

Table 1 Sample field sites of wheat and corn

Field	Crop name	Location	Zero-sites *N*	Sites-*N*
WH	Wheat	45°29′32″ N	WH—01	WH 02–26
		80°54′22″ W	WH—36	WH 18–50
CN	Corn	43°27′40″ N	CN—01	CN 02–07
		80°48′53″ W	CN—36	CN 01–11

3.3 Field Sites Sampling

The site fields were in Easthope Township, Ontario, in corn and wheat two fields 90% of the land in this region will be used for agricultural production, making it a very productive agricultural area. The area's soils are predominantly clay and silty loams with adequate natural drainage. Summers are moderate to hot (average air temperature 20.2 °C), and the climate is tropical continental. The two principal annual crops in the research region are winter Triticum aestivum (wheat) and Zea mays (corn). Usually, corn is planted in May and yields in early September or November. During the winter, it goes dormant and then blooms another time from late March to early April, just during time for harvest. Except for the control area in WH and CN in N sites fields, the Ministry of Agriculture, Food and Rural Affairs advised a rate of fertiliser treatment of 105 kg for winter wheat and 134 kg for corn for the majority of the fields, as shown in Table 1. The data for wheat and maize were obtained between May and September, according to the growing seasons of the two crops. The majority of the research locations were at least 35 m apart, and each location was within a distinct Landsat satellite pixel. There are also individual sampling locations and no-nitrogen plots available (Fig. 5).

3.4 Field Collection Data

At each sampling site, it acquires representative of five leaves from the higher plant canopy for and leaf reflectance measurements and biochemical analysis. I placed leaf models in synthetic bags and maintained in the dark at a temperature (0 °C). The samples are after that promptly sent support to the University of Toronto, where they were procedure within 4 h after being collected. Spectrophotometric grade N, N-dimethyl form amide was used to extract foliar chlorophyll, and it measured absorbance at 662.9, 644.9 and 482 nm using a Shimadzu spectrophotometer (UV-170). Each sampling site's measured chlorophyll content values was determined as

Fig. 5 Sampling filed sites for wheat and corn

mean values from five leaf models are obtained and it measured leaf chlorophyll content on each sampling date.

An Analytical of Spectral Devices was used to measure the leaf reflectance, nitrogen transmittance and temperature (SD) to eliminate the influence of electrical noise; they got a dark dimension previous to each leaf reflectance example. The reflectance spectrum R_{μ} of each leaf was determined because the leaf index area (LIA) measured by placing a leaf in the model port of the integrated sphere. By placing a leaf in the sampling port of the integrating sphere and computing the spectrum reflectance R of each leaf, it assessed LIA at the same sampling places on the same days as leaf sampling.

$$\text{LIA } R_{\mu} = \left[\frac{\text{Leaf Index}_{\mu}}{\text{Leaf panel radiation}_{\mu}} \right] \tag{4}$$

3.5 LLR Scaling Ratio of H and UH (Healthy and Unhealthy)

SAIL's leaf reflectance input has a big impact on the output of the predicted reflectance canopy. Using a strong leaf reflectance to create the LLR instead of a familiar model, canopy reflectance could result in a bias in the leaf-level inversion for stressed leaves. It developed healthy and unhealthy leaf LLRs to adjust for this potential bias, and pixel applied inverse distance weighting pixel for the finest match. In order to the LIA conditions and pixel's solar zenith angle, scaling ratio (μ) value is calculated by comparing each pixel in the Landsat image to the two LLRs (H and UH) value. An inverse distance weighting—IDW—method was utilised to establish a reasonable ratio value between the healthy and unhealthy LLR matches. They found IDW on the idea that points that are closer together are more similar than

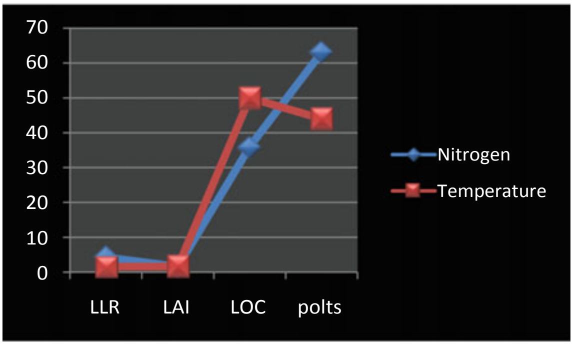


Fig. 6 Leaf chlorophyll maps for *N&T*

points that are farther apart. According to this idea, if a Landsat pixel's reflectance canopy be extra close in the direction of the familiar simulated reflectance canopy by means of the unhealthy leaf reflectance, the scaling ratio (μ) values used to calculate leaf reflectance would have a higher influence. They compared the LLR healthy and unhealthy matching to the reflectance canopy from the Landsat scene. It added a weighting to the ratio based on the isolation (difference) among the Landsat picture of healthy and unhealthy. Equation (5) produces the LLR scaling ratio of H and UH (Fig. 6).

$$\mathrm{LLR}(\mu) = \frac{\left(\frac{\mathrm{LLR}_{\mathrm{healthy}}(\mu)}{R_{\mathrm{healthy}}} + \frac{\mathrm{LLR}_{\mathrm{unhealthy}}(\mu)}{R_{\mathrm{unhealthy}}}\right)}{\left(\frac{1}{R_{\mathrm{healthy}}} + \frac{1}{R_{\mathrm{unhealthy}}}\right)} \tag{5}$$

$$\begin{aligned} \mathrm{LLR}_{\mathrm{healthy}} &= \left[\mathrm{Sample}\,R_{\mathrm{healthy}\ (\mu)} - \mathrm{Spad}\,R_{\mathrm{healthy}(\mu)}\right]^2 \\ \mathrm{LLR}_{\mathrm{unhealthy}} &= \left[\mathrm{Sample}\,R_{\mathrm{unhealthy}(\mu)} - \mathrm{Spad}\,R_{\mathrm{unhealthy}(\mu)}\right]^2 \end{aligned} \tag{6}$$

4 Result and Discussion

Wheat and corn *N* fertilised sites' temporal trends calculated leaf chlorophyll content of nitrogen as well as temperature following sowing over the previous winter, and winter wheat begins to grow earlier DOY 130. In the midst of the nitrogen fertiliser, growing season application had a significant collision on peak values of mutually LLR with chlorophyll. Fertilised wheat grasslands have a highest average (4.3) of LLR, although wheat non-fertilised fields no more than had a highest average LAI (1.6). A photosynthetic activity in chlorophyll fluorescence was used to LCC assess photosynthetic activities in Swiss leaf chlorophyll content approximate using the SPAD-503 (Map) value of leaf by different treatments medium. Despite the fact that nitrogen was applied to the sites with lower zero-*N* values in fertilizer. Changes in leaf chlorophyll content reveal disparities between nitrogen and no-nitrogen treatment locations during the growth season; with fertilised sites, achieving higher leaf chlorophyll content is higher than in non-fertilised spots. The maximal chlorophyll concentration in wheat fertilised plots is 63 g cm^2, but it is only 44 g cm^2 in non-fertilised plots. The chlorophyll content of corn is affected by nitrogen and temperature application to a higher extent than LLR values, with maximum values of 83 g cm^2 and 35 g cm^2 for N and non-zero sites, respectively (Table 2 and Fig. 7).

Table 2 Numerical results of SPAD units value (5 map)

Spectral indices	Corn	Wheat	Corn and wheat
LIA	0.38	0.32	0.24
LCC	0.49	0.39	0.34
NDVI	0.55	0.44	0.37
SI	0.56	0.48	0.46
RDVI	0.47	0.45	0.44
SPAD map 3	0.56	0.48	0.56
SPAD map 4	0.6	0.49	0.49
SPAD map 5	0.8	0.54	0.59

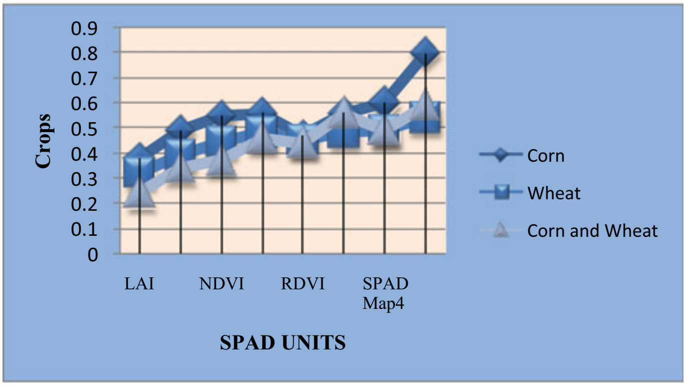
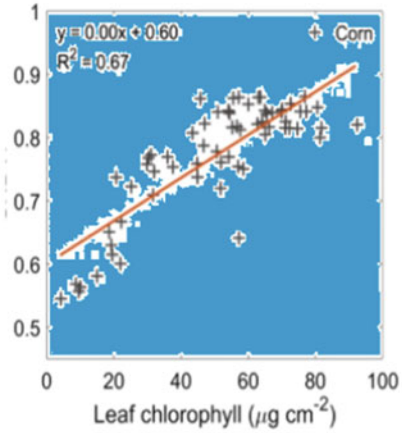
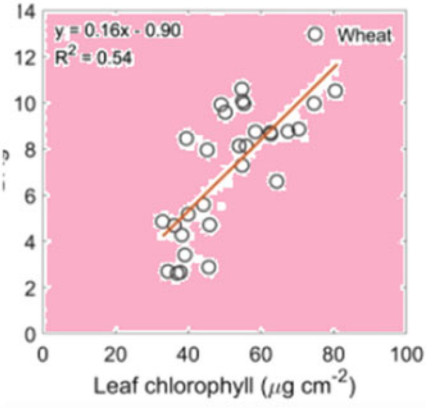


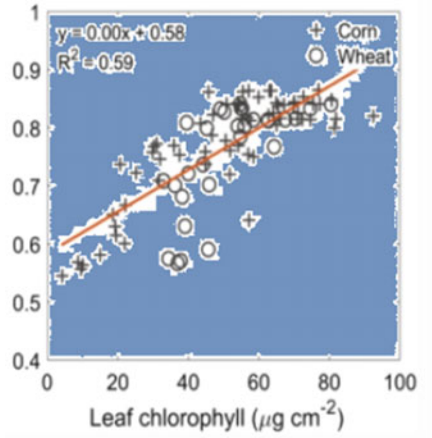
Fig. 7 Numerical results for crops and SPAD



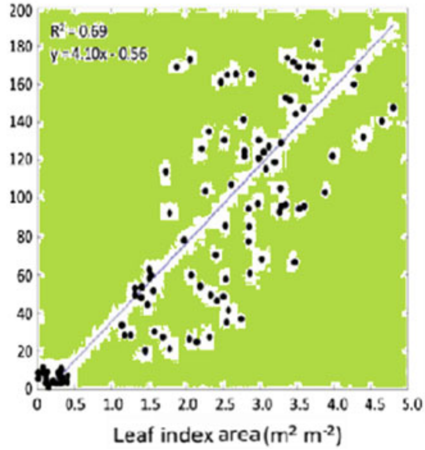
(a)



(b)



(C)



(d)

5 Conclusions

The SPAD-503 chlorophyll metre was identified as a viable device for assisting in the generation level of chlorophyll and yield plans for recognising geographic differences and serving as a standard for crop health monitoring in precision agriculture in this paper. A chlorophyll metre was used to help determine N stress levels so that more accurate N control could have implemented. The intricacy of stress-related factors as measured by chlorophyll gauge difficult to isolate the chlorophyll metre could use as a supplement to other crop sensing and nitrogen delivery strategies. Variable-rate application systems can be developed using the database of chlorophyll level readings with yield plans, and soil fertility and chlorophyll level maps can assist in geographic differences for site-specific cultivation operations.

References

1. Croft H, Chen JM, Zhang Y, Simic A (2013) Modelling leaf chlorophyll content in broadleaf and needle leaf canopies from ground, CASI, Landsat TM 5 and MERIS reflectance data. Remote Sens Environ 133:128–140
2. Croft H, Chen JM (2018) Leaf pigment content. In Liang S (ed) Comprehensive remote sensing, Vol 3, pp 117–142
3. Croft H, Chen JM, Luo X, Bartlett P, Chen B, Staebler RM (2017) Leaf chlorophyll content as a proxy for leaf photosynthetic capacity. Glob Change Biol 23(9):3513–3524
4. Darvishzadeh R, Skidmore A, Abdullah H, Cherenet E, Ali A, Wang T, Nieuwenhuis W, Heurich M, Vrieling A, O'Connor B, Paganini M (2019) Mapping leaf chlorophyll content from Sentinel-2 and RapidEye data in spruce stands using the invertible forest reflectance model. Int J Appl Earth Obs Geoinf 79:58–70
5. Dong T, Liu J, Qian B, Jing Q, Croft H, Chen J, Wang J, Huffman T, Shang J, Chen P (2016) Deriving maximum light use efficiency from crop growth model and satellite data to improve crop biomass estimation. IEEE J Sel Top Appl Earth Observ Remote Sens 10(1):104–117
6. Croft H, Arabian J, Chen JM, Shang J, Liu J (2020) Mapping within-field leaf chlorophyll content in agricultural crops for nitrogen management using Landsat-8 imagery. Precision Agric 21(4):856–880
7. Vinoth Kumar CNS, Suhasini A (2017) Springer book chapter entitled "secured three-tier architecture for wireless sensor networks using chaotic neural networks". In: Advances in intelligent systems and computing AISC Series, Springer Science + Business Media Singapore 2017, Vol 507, Chapter 13, pp 129–136, ISSN 2194-5357. doi:https://doi.org/10.1007/978-981-10-2471-9_13
8. Rathi R, Balyan N, Vinoth Kumar CNS (2020) Pneumonia detection using chest X-Ray. Int J Pharma Res (IJPR) 12(3). ISSN: 0975-2366. https://doi.org/10.31838/ijpr/2020.12.03.181
9. Baranidharan B, Vinoth Kumar CNS, Babu MV (2021) An improved inception layer-based convolutional neural network for identifying rice leaf diseases. In: Intelligent learning for computer vision. CIS 2020, lecture notes on data engineering and communications technologies (LNDECT), Vol 61. Springer, Singapore. ISSN 2367-4512, pp 119–129. https://doi.org/10.1007/978-981-33-4582-9_10
10. Venkatesh K, Parthiban S, Kumar PS, Vinoth Kumar CNS (2021) IoT based unified approach for women safety alert using GSM. In: Proceedings of the 3rd international conference on intelligent communication technologies and virtual mobile networks, ICICV 2021, pp 388–392, 9388415
11. Luo X, Croft H, Chen JM, He L, Keenan TF (2019) Improved estimates of global terrestrial photosynthesis using information on leaf chlorophyll content. Glob Change Biol 25(7):2499–2514
12. Sun Q, Jiao Q, Qian X, Liu L, Liu X, Dai H (2021) Improving the retrieval of crop canopy chlorophyll content using vegetation index combinations. Remote Sens 13(3):470
13. Qian X, Liu L (2020) Retrieving crop leaf chlorophyll content using an improved look-up-table approach by combining multiple canopy structures and soil backgrounds. Remote Sens 12(13):2139
14. Zhang K, Liu X, Ma Y, Zhang R, Cao Q, Zhu Y, Cao W, Tian Y (2020) A comparative assessment of measures of leaf nitrogen in rice using two leaf-clip meters. Sensors 20(1):175

S. Sakthipriya completed her Master of Engineering in the field of Computer Science and Engineering in ARM College of Engineering and Technology Affiliated to Anna University Chennai in the year 2019. She completed her Bachelor of Engineering in Alpha College of Engineering and Technology Affiliated to Anna University Chennai in the year 2013. She is presently as Research Scholar at the Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India. Her main areas of research are precision agriculture in machine learning and IoT.

Dr. R. Naresh completed his Ph.D. in Computer Science and Engineering in Anna University Chennai in the year 2017. He completed Master of Engineering in the field of Computer Science and Engineering in Anna University Chennai in the year 2011. He completed his Bachelor of Engineering under Anna University Chennai in the year 2007. He was working as Assistant Professor at Anna University Chennai (University College of Engineering Tindivanam), Chennai, India, from 2011 to 2018. He is presently working as Associate Professor at the Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India. His main thrust research areas are group key management in network security and cloud security.

A Decision Support System for Understanding the Importance of Two-Dosage Vaccination

R. Saraswathi Meena, R. Saranya, C. Deisy, S. Parthasarathy, S. Sridevi, and M. K. A. A. Khan

Abstract Nowadays, vaccination plays the major role in controlling the death rates in COVID-19. However, certain people don't have a trust on vaccines like Covaxin and Covishield. In order to make the decision on these vaccines like Covaxin and Covishield, decision support system (DSS) is necessary using machine learning approach. Many researchers find the supervised learning model gives best prediction among many machine learning models for COVID-19 situation. Supervised learning model has regression and classification models. Here, supervised machine learning models like logistic regression, decision tree and random forest are used to create a decision support system (DSS). With the use of this machine learning models, people can understand the uses of COVID-19 vaccines.

Keywords Machine learning · Logistic regression · Decision tree · Random forest

1 Introduction

Coronavirus disease is an infectious disease caused by a Coronavirus. People affected by the Coronavirus have mild to severe respiratory problems. People have medical

R. S. Meena (✉) · R. Saranya · S. Parthasarathy
Department of Applied Mathematics and Computational Science, Thiagarajar College of Engineering, Madurai, India
e-mail: saraswathimeenacse@gmail.com

R. Saranya
e-mail: saranyarajagopal2002@gmail.com

C. Deisy · S. Sridevi
Department Information Technology, Thiagarajar College of Engineering, Madurai, India
e-mail: cdcse@tce.edu

S. Sridevi
e-mail: sridevi@tce.edu

M. K. A. A. Khan
Faculty of Engineering, UCSI University, Selayang, Malaysia

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_23

problems like respiratory diseases, diabetes, pressure and cardiovascular diseases are leading to severe illness/deaths. For this coronavirus, World Health Organization has administrated 13 vaccines. In India, most of them use the Covishield and Covaxin vaccines to prevent the COVID virus. However, some peoples do not have trust in these vaccines. If people take the vaccines approved by World Health Organization (WHO), it will improve our immunity systems against COVID-19. In order to improve the trust among peoples, we use a machine learning algorithm in this paper.

Machine learning is to allow the computers learn automatically using data and algorithms. Machine learning contains supervised learning, unsupervised learning and reinforcement learning. Many researchers found that supervised learning is best for the COVID-19 among machine learning algorithms. Supervised learning is that the sorts of machine learning during which machines are trained using well-labelled training data. We use this case labelled data set for prediction. The vaccinated peoples count details and death date details are used for this prediction. Here supervised learning algorithm is used to train the machines. Here, some of the supervised machine learning, like decision tree, random forest and logistic regressions are used to train the machines. We predict the correlation between vaccinated people and death details through these supervised learning algorithms.

If the vaccinated people count increases, then death rate decreases, peoples can understand the power of vaccination. In this case, we take three supervised learning algorithms to continue this proof. The supervised learning algorithms have classification and regression algorithms. In these types, from classification, random forest and decision tree are used for prediction of death rates based on vaccinated people count. From regression, we used logistic regression for predicting the death rates based on the vaccinated people count.

Based on the predictions of all three algorithms, the relationship between death rates and vaccinated people count is predicted. Through this result, people can find the effectiveness of vaccination. Without confusion, people can be vaccinated through this machine learning algorithm prediction on the relationship between death rates and vaccinated people count.

2 Related Researches

Kwekha-Rashid et al. [1] have used supervised learning, which gives the best result than unsupervised model in machine learning algorithms for the health centre programmes and COVID-19 cases.

Chaudhary et al. [2] have used the principal component analysis on the COVID-19 data set for dimensionality reduction and fins most vital variables. Further, this paper used unsupervised algorithms like K-Means to find the community structure of countries.

Aggarwal et al. [5] in this paper, an additive utility assumption-based approach for multi-criterion decision support system validated empirically using the susceptible, exposed, infected and recovered. The results include comparative analysis has the parameters such as Precision, Recall and F1-Score.

Rasheed et al. [8] during this paper for the automated diagnosis of Coronavirus with high accuracy from X-ray images, two classifiers are used. Those two classifiers are logistic regression and convolutional neural networks. In order to achieve the system, they used efficient and fast principal component analysis for the dimensionality reduction. Therefore, this paper uses PCA for the dimensionality reduction and logistics regression for the data modelling.

Jawahar et al. [9] this study aims to research the chest X-ray for determining the presence of COVID-19 employing a machine learning algorithm. These researchers use a various methods of machine learning algorithms and deep learning approaches to predict COVID-19. However, getting an accurate solution using these AI techniques is that the major challenge remains hospitable researchers.

Shuai et al. [10] this paper uses StandardScaler normalization method to preprocess the data and use K-Means clustering, Agglomerative clustering and density-based spatial clustering of applications with noise (DBSCAN) clustering algorithm to perform cluster analysis, use Silhouette Coefficient index, Calinski–Harabasz score, Davies–Bouldin score to evaluate the clustering results, provide data support and strategic support for the prevention and control of the epidemic almost like COVID-19 within the future.

Rustam et al. [13] this study exhibits the potential of ML models towards forecasting the amount of upcoming patients affected through COVID-19. In order to model the data set forecasting, four machine learning algorithms used like linear regression (LR), least absolute shrinkage and selection operator (LASSO), support vector machine (SVM) and exponential smoothing (ES).

Khanday et al. [14] have worked on clinical data set to analyse, and the clinical data set classified into four different classes like term frequency/inverse document frequency (TF/IDF), bag of words (BOW) and report length. Then they used some machine learning algorithms like logistic regression, multinomial Naïve Bayes, for COVID-19 prediction.

Pourhomayoun et al. [15] have chosen some machine learning algorithms like support vector machine (SVM), artificial neural networks, random forest, decision tree, logistic regression and K-nearest neighbour (KNN) algorithms to find out the death rate caused by COVID-19 and uses confusion matrix for finding sensitivity.

Linda Wang et al. [16] have analysed the chest X-ray images to find abnormalities related to analysing COVID-19. They have created their own architecture for the chest X-ray images with the use of machine learning.

Pal et al. [17] have used the neural network for country wise COVID-19 prediction. They take every country specific data set for the observation through that they achieve accuracy. They used long short-term memory (LTSM) and they compared with the recurrent neural network (RNN) and gated recurrent units (GRU). So that they have proposed the optimized long short-term memory systems to predict the COVID-19 country wise with the use of neural networks.

3 Proposed System

In this work, the vaccinated people information is given to the system as input. Based on the vaccinated people count and death count, the system finds the death rate. If the death rate decreases, the vaccination plays the major role in the death cases caused by the COVID-19 cases. For the prediction of death rate, in this system, machine learning algorithms are used. Supervised machine learning algorithms give the accurate death rate when compared to the unsupervised learning algorithms. So this system takes the supervised machine learning algorithms, for the prediction of death rate caused by the COVID-19.

Three supervised learning algorithms are used in this system to forecast death rates. The supervised learning methods for death rate prediction are random forest, logistic regression and decision tree. If the death rate falls, our system will recommend the vaccine to people who are unaware of its benefits. Although it will discover the optimal learning method for death rate prediction based on mean squared error, mean absolute error, Precision, Recall and F1-Score among the three supervised machine learning techniques listed above.

4 Implementation

In this section, using data from various sources preprocessing is implemented. After preprocessing, visualization on various features displays and three different models: logistic regression, random forest and decision tree are trained on the data set. Finally, model interpretation is implemented.

4.1 Data Description and Data Preprocessing

We imported data from various sources, like Kaggle, which are available in the format .csv and .xlsx. The raw data set comprises various attributes like the number of person's vaccinated first dose and the second dose at the same time number of patients got affected, cured and dead on that day to the disease COVID-19. Data sets in any data processing applications can have missing data values.

These missing values can affect the parameters and model while predicting values. Hence, it should be removed. Selecting the needed attribute for training a model is part of preprocessing. It is because attributes play a major role to train a model in predicting. There are numerous columns, but only 5 to 7 have a significant role in vaccination decision-making. Along with this, calculating the required attributes like discharge rate, active cases and death rate.

$$\text{Active Cases} = \text{No of conformed cases}$$

$$\begin{aligned}
&- (\text{no of curved cases} + \text{no of death cases})\\
\text{Discharge Rate} &= (\text{No of curved cases till date}\\
&/\text{No of confirmed cases}) * 100\\
\text{Death Rate} &= (\text{No of death till date}\\
&/\text{No of confirmed cases till date}) * 100
\end{aligned}$$

Normalization of data is the last step in preprocessing. The rationale behind the normalization is to possess data on a standard scale without distorting differences within the ranges of values. Normalization is rescaling the values in the range 0–1 in numeric attributes. In this technique, we implement linear transformation in the data. The below formula is for Min–max scalar finding.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

During preprocessing and collecting, the data set contains some missing values. Thus, NA/N values are converted to zero or they may be removed from the data set.

4.2 Feature Selection

We implemented the technique feature selection when the number of attributes in the data set is higher. Therefore, many attributes are unimportant for prediction and may lead to lesser accuracy. The main benefit of employing feature selection is that it reduces over-fitting while allowing for less noise in decision-making. When the number of attributes is smaller than the time it takes to train, the method will be smaller as well.

Here, we use Person's Correlation Coefficient as a feature selection algorithm. When two features are highly correlated then they are linearly dependent and hence have almost an equivalent effect on the dependent variable. The range of correlation is ±r. (Here, +r means highly correlated) in case of −r then they are uncorrelated. If the correlation is 0, then values are independent. Usage of correlation coefficient reduces the dimensionality and reduces the added noise that affects the predicted attribute. To find a person's correlation coefficient the below formula is used.

$$x' = \frac{\sum((X_i - X')(Y_i - Y')}{\sum((X_i - X')(Y_i - Y'))} \tag{2}$$

As a result, the attributes that are highly connected with the prediction attribute are found using Person's Correlation method. Choosing the property with a correlation value (r) greater than 0.7 to select the aspects that are significantly connected (Fig. 1 and Table 1). (We excluded cured and death, as well as a few other attributes, because the death and discharge rates are known.)

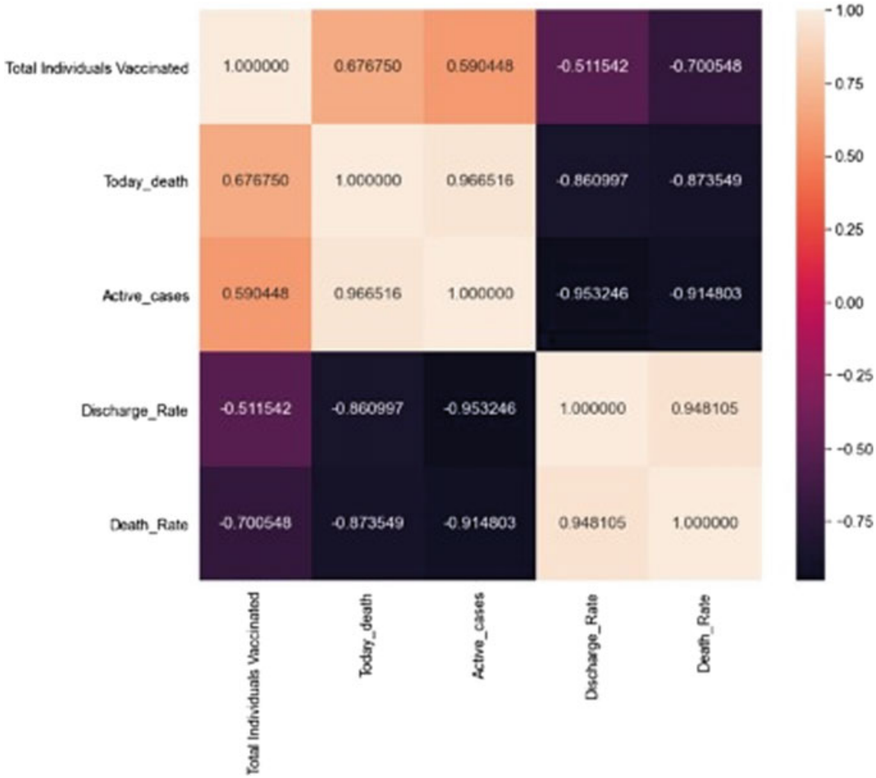


Fig. 1 Correlation plot for various attribute

Table 1 Result of various methods

Attribute with respect to death rate	Person's correlation value
Total individuals vaccinated	0.70055
Today death	0.87355
Active cases	0.91480
Discharge rate	0.94810

4.3 Machine Learning Models

Machine learning models are used to recognize a pattern by training different data sets. Train a model over a group of knowledge, providing it with an algorithm that it can use to reason over and learn from those data.

Decision Tree: We mostly used decision tree in classification and regression. Decision tree builds a model in the tree's form structure. Data set breaks into smaller and smaller subsets, but at an equivalent duration, an association of decision trees has been developed. The final decision tree comprises a root node, the decision nodes and leaf nodes. The decision tree tries to find decision rules that make the best split measured by Gini impurity or entropy.

Random Forest: Random forest selects a random sample from the data set, such as the algorithm will construct a decision tree for every sample. Finally, considering the class that is in the mode of the mean predictions of all the outputs from each decision tree. Random forest produces greater results without using hyper-parameter tuning. Therefore, it is a much more widely accepted algorithm for its performance and simplicity. It is difficult for humans to interpret many decision trees; therefore, the model is small amount of interpretable than a single decision tree.

Logistic Regression: Logistic regression is a machine learning algorithm that takes real-valued inputs and predicts the probability of the target variable. Therefore, the dependent value is binary natured variable with (1/0) or (yes/no).

4.4 Training the Model

In terms of implementation, the data is divided into two groups: train and test (55 rows for training and remaining data for testing). There is no hyper-parameter when evaluating the decision tree and random forest tree models. In random forest tree, we initialized the random state as an integer.

$$P(Y) = \frac{1}{(1 + e\hat{}(\beta 0 + \beta 1x1 + \beta 2x2 + \beta 3x3 + \beta 4x4))} \tag{3}$$

In order to solve the multi-class problem in logistic regression, "lbfgs" solver is used to handle multinomial loss. Finally, all these algorithms and method are implemented using scikit-learn and Python 3.8.5. After training, the model is being tested using test data, and the accuracy will be determined (Figs. 2, 3 and 4).

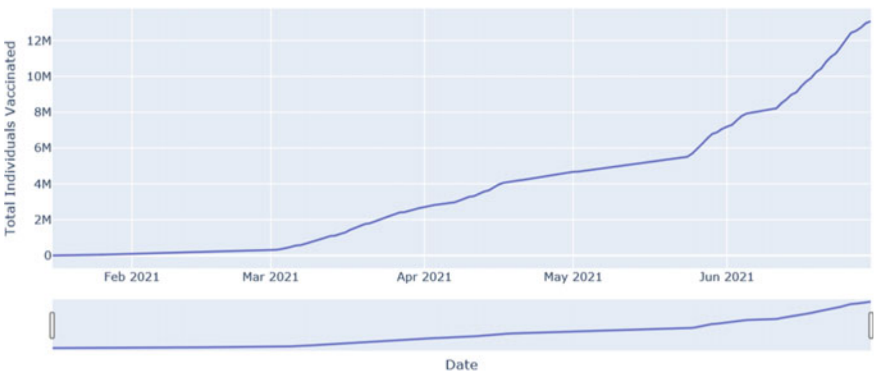


Fig. 2 Total individual vaccinated from day to day

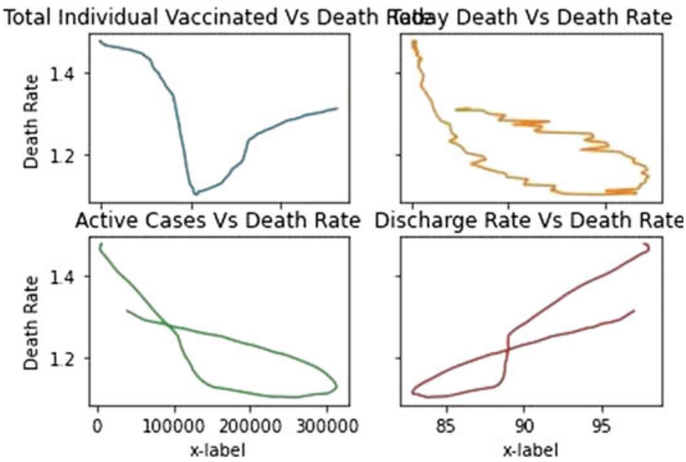


Fig. 3 Density plot for death rate with respect to various attribute

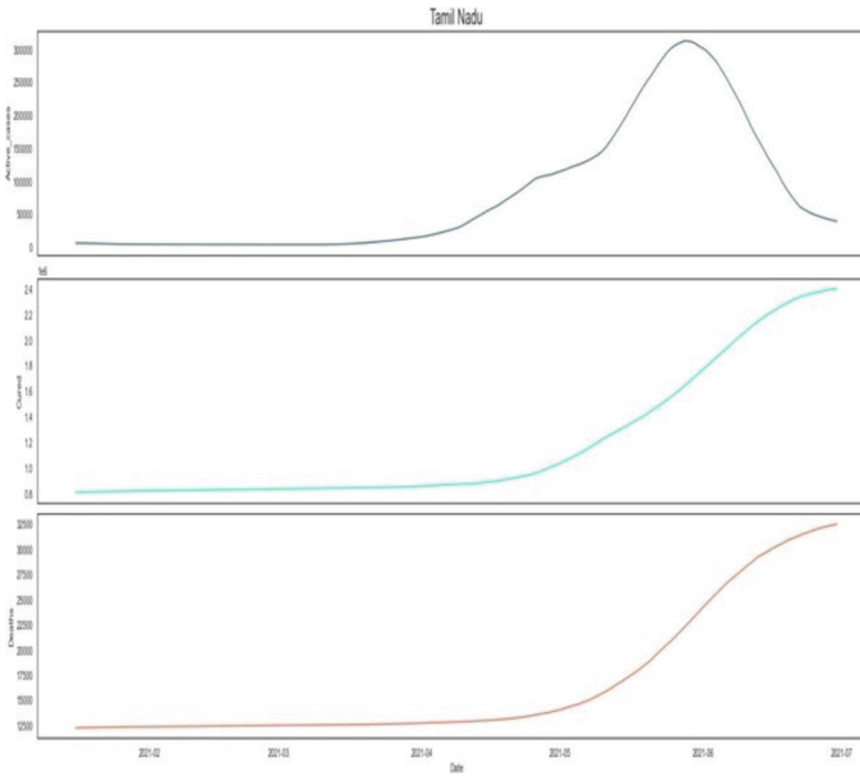


Fig. 4 Density graph on important factors from day to day

4.5 Performance from the Model

The efficiency of the algorithms cannot be determined only from accuracy. Some more metrics are used to measure the efficiency of the trained models. Some standard performers like Precision, Recall and mean absolute error are used to classify the models. Various graphs compare the performance of various models. Accuracy is the percentage of correct predictions with total observation implemented in the test data. We define accuracy as the most intuitive performance measure.

Even though the accuracy is higher, but the model failed to make the best fit so, the model fails. Hence, we calculate mean squared error (MSE), root mean squared error (RMSE) and mean absolute error (MAE) to conclude whether the model is fit or not.

$$\text{MSE} = \frac{\sum((Y_i - Y_i')^2)}{n} \tag{4}$$

Mean absolute error (MAE) is the sum of errors between the predicted and the actual value in the absolute form.

$$\text{MAE} = \frac{\sum\left|(Y_i - Y_i')\right|}{n} \tag{5}$$

Mean squared error (MSE) is calculated by taking the typical of the square of the difference between the particular and predicted value. Root mean squared error (RMSE) is calculated by taking square root for the term mean squared error (MSE) value.

$$\text{RMSE} = \frac{\sqrt{\Sigma((Y_i - Y_i')^2)}}{\sqrt{n}} \tag{6}$$

Precision is calculated as the ratio of True Positive (TP) to the Total Predicted Positive observations. However, Recall is calculated as the ratio of True Positive to the Total Positive in the actual class. F1-Score defines the balance between Precision and Recall. Precision (P) is useful when the False Positive (FP) is high. However, Recall (R) works when False Negative (FN) is high. Finally, F1-Score gives a better measure of the incorrectly classified cases compared to accuracy metrics.

$$\text{F1-Score} = \frac{(2 * P * R)}{(P + R)} \tag{7}$$

Table 2 Result of various methods

Algorithm	Accuracy (%)	MSE	RMSE	MAE	Precision (%)	Recall (%)	F1-score (%)
Random forest	91.67	0.078	0.279	0.078	91.43	94.12	92.75
Decision tree	87.93	1.029	1.014	1.029	90.91	88.24	89.55
Logistic regression	93.33	0.032	0.178	0.031	94.12	94.12	94.12

5 Result

COVID-19 is the current pandemic disease that challenges the medical field. Here we conclude that the death rate decreases as the number of individuals vaccinated increases. From this prediction, the person vaccinated will have a very less death rate. We used decision tree, random forest and logistic regression models for predictions. Thus, we conclude that logistic regression produces better accuracy (Table 2).

6 Conclusion

This system proposed the solution to recommend the vaccine for the people who unaware of the importance and benefits of vaccination. For that, this system uses the information about vaccinated people count and death rate count as an input. For the prediction of death rate, it uses three machine learning algorithms such as random forest, logistic regression and decision tree. These algorithms find that if vaccinated people count increases automatically it will decreases the death rate. Therefore, this system recommends the people to take vaccine in order to avoid the COVID-19. Although based on the accuracy of these three algorithms, it will find that logistics regression gives the best result for death rate prediction in COVID-19 cases based on the accuracy.

References

1. Kwekha-Rashid AS, Abduljabbar HN, Alhayani B (2021) Coronavirus disease (COVID-19) cases analysis using machine-learning applications. Appl Nanosci 1–13
2. Chaudhary L, Singh B (2021) Community detection using unsupervised machine learning techniques on COVID-19 dataset. Soc Netw Anal Min 11(1):1–9
3. Vardhini PH, Prasad SS, Korra SN (2021) Medicine allotment for COVID-19 patients by statistical data analysis. In 2021 international conference on emerging smart computing and informatics (ESCI). IEEE, pp 665–669

4. Vega C (2021) From Hume to Wuhan: an epistemological journey on the problem of induction in COVID-19 machine learning models and its impact upon medical research. IEEE Access 9:97243–97250
5. Aggarwal L, Goswami P, Sachdeva S (2021) Multi-criterion intelligent decision support system for COVID-19. Appl Soft Comput 101:107056
6. Medina J, Rojas-Cessa R, Umpaichitra V (2021) Reducing COVID-19 cases and deaths by applying blockchain in vaccination rollout management. IEEE Open J Eng Med Biol
7. Zoabi Y, Deri-Rozov S, Shomron N (2021) Machine learning-based prediction of COVID-19 diagnosis based on symptoms. NPJ Dig Med 4(1):1–5
8. Rasheed J, Hameed AA, Djeddi C, Jamil A, Al-Turjman F (2021) A machine learning-based framework for diagnosis of COVID-19 from chest X-ray images. Interdisc Sci Comput Life Sci 13(1):103–117
9. Jawahar M, Anbarasi LJ, Prassanna J, Jackson CJ, Manikandan R, Alzubi JA, Dhanya D (2021) Diagnosis of covid-19 using optimized pca based local binary pattern features. Int J Curr Res Rev 13(6 Special Issue)
10. Shuai Y, Jiang C, Su X, Yuan C, Huang X (2020) A hybrid clustering model for analyzing COVID-19 national prevention and control strategy. In 2020 IEEE 6th international conference on control science and systems engineering (ICCSSE). IEEE, pp 68–71
11. Sonbhadra SK, Agarwal S, Nagabhushan P (2020) Target specific mining of COVID-19 scholarly articles using one-class approach. Chaos, Solitons Fractals 140:110155
12. Anandhanathan P, Gopalan P (2021) Comparison of machine learning algorithm for COVID-19 death risk prediction
13. Rustam F, Reshi AA, Mehmood A, Ullah S, On BW, Aslam W, Choi GS (2020) COVID-19 future forecasting using supervised machine learning models. IEEE Access 8:101489–101499
14. Khanday AMUD, Rabani ST, Khan QR, Rouf N, Din MMU (2020) Machine learning based approaches for detecting COVID-19 using clinical text data. Int J Inf Technol 12(3):731–739
15. Pourhomayoun M, Shakibi M (2020) Predicting mortality risk in patients with COVID-19 using artificial intelligence to help medical decision-making. MedRxiv
16. Wang L, Wong A, COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. arXiv:2003.09871
17. Pal R, Sk AA, Kar S, Prasad D (2020) Neural network based country wise risk prediction of COVID-19. https://doi.org/10.20944/preprints202004.0421.v1
18. Olsen W, Bera M, Dubey A, Kim J, Wiśniowski A, Yadav P (2020) Hierarchical modelling of COVID-19 death risk in India in the early phase of the pandemic. Europ J Develop Res 32(5):1476–1503
19. Wu H, Ruan W, Wang J, Zheng D, Liu B, Geng Y, Helal S (2021) Interpretable machine learning for covid-19: an empirical study on severity prediction task. IEEE Trans Artif Intell
20. Iwendi C, Bashir AK, Peshkar A, Sujatha R, Chatterjee JM, Pasupuleti S, Jo O (2020) COVID-19 patient health prediction using boosted random forest algorithm. Front Public Health 8. https://doi.org/10.3389/fpubh.2020.00357
21. Sujath R, Chatterjee JM, Hassanien AE (2020) A machine learning forecasting model for COVID-19 pandemic in India. Stoch Env Res Risk Assess. https://doi.org/10.1007/s00477-02001827-8
22. Bhanu K (2020) Analysis, prediction and evaluation of COVID-19 datasets using machine learning algorithms. Int J Emerg Trends Eng Res 8. https://doi.org/10.30534/ijeter/2020/117852020
23. Vijayashree J, Sriman Narayana INC (2016) Heart disease prediction system using data mining and hybrid intelligent techniques: a review. Int J Bio-Sci BioTechnol 8:139–148. https://doi.org/10.14257/ijbsbt.2016.8.4.16
24. Panda D, Ray R, Abdullah A, Dash S (2019) Predictive systems: role of feature selection in prediction of heart disease. J Phys: Conf Ser 1372:012074. https://doi.org/10.1088/1742-6596/1372/1/012074

Sanskriti—A Distributed E-Commerce Site Implementation Using BlockChain

Soumitro Datta, Kunal Singh Naruka, S. A. Sajidha, V. M. Nisha, and Ramesh Ragala

Abstract Application for decentralised blockchain adoption has no end. Obstacle for deployments revolves around recognition, willingness to take the lead and overcoming a privacy risk obstacle. Our research work is an implementation of this application in the e-commerce sector. In the real-world scenario of e-commerce, small businesses have to face the adverse effects of centralised authorities and their monopoly, and it becomes really hard for them to register or advertise their products on centralised platforms. Security and privacy are also huge issues in case of centralised platforms. Therefore, we propose a decentralised e-commerce platform for fast KYC system which can be completed in simple and easy steps and can be done from any convenient place. We use a smart contract system which consists of a set of rules that parties in a smart contract agree to follow. Our proposed approach is both quick and secure that stores data in parts into MongoDB server and blockchain. This also improves user privacy. In case of any malicious entry into the server, the attacker will have his hands only on some and irrelevant data which cannot be decoded. Making this division will also reduce the cost of writing all data on blockchain. Along with these, the implementation supports INR payments as well for ease of access.

Keywords E-commerce · Cryptocurrency · Decentralisation · Security · Blockchain · Ethereum

S. Datta (✉) · K. S. Naruka · S. A. Sajidha · V. M. Nisha · R. Ragala
School of Computer Science and Engineering, VIT University, Chennai, Tamil Nadu, India
e-mail: soumitro.datta0705@gmail.com

S. A. Sajidha
e-mail: sajidha.sa@vit.ac.in

V. M. Nisha
e-mail: nishavm@vit.ac.in

R. Ragala
e-mail: ramesh.ragala@vit.ac.in

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_24

1 Introduction

Blockchain is a new software and development platform that is still in its infancy. It is a new paradigm that aims to empower peers while also removing the need for a central authentication authority. Blockchain has effectively implemented a new level of distributed sovereignty. Cryptocurrency is a well-known result of this new computing paradigm, which has gotten a lot of attention. Purpose of blockchain can be widely seen as the foundation of Bitcoin e-currency. Recently, many other applications can be seen in various fields.

The most well-known application of blockchain is in the cryptocurrency industry. Blockchain can improve supply chain efficiency and transparency. Blockchains allow for the transfer of funds anywhere in the world, which is ideal for a globalised supply chain. Normally, an e-commerce corporation saves this information in a centralised place or on the cloud, and it is always vulnerable to hackers. Blockchain saves data on a decentralised system with fewer hacking opportunities.

It has the potential to change the way businesses operate and can aid in the security and efficiency of businesses.

E-commerce platforms house a lot of information. Customer information is stored on centralised servers, making it vulnerable to cybercriminals. Marketing field is not often transparent making it difficult for small companies to establish themselves. Bringing in blockchain integrated solution here can help solve these issues. Buyers and sellers are connected directly through a peer-to-peer network without the participation of middle-men in our proposed system. Triad of producers, vendors, and purchasers are represented on our decentralised market. This technology can be used by e-commerce businesses to store consumer information [1].

Another major difficulty for centralised e-commerce firms is ethical conduct [2, 3]. It is no secret that transparency is becoming a hot topic among investors and government organisations alike. The United States President Donald J. Trump expressed his concern about the transparency of e-commerce giants like Amazon in a recent Twitter post. Any modern technology, providing safety takes time for users to trust it. Most individuals are unaware of how much information they unwillingly give up. People may regain control of their data by using blockchain-based privacy solutions. E-commerce can be improved by implementing blockchain technology, in which all transactions are kept in a shared ledger (smart contract) and are almost difficult to change (immutable) by anybody [4]. For customers, it will ensure that the online platform infrastructure remains intact.

Our proposed a solution, which divides and stores information on blockchain and MongoDB separately, makes the platform even more secure and also improves on privacy. Hence, in case of a malicious entry into the server, the attacker will have his hands only on some and irrelevant data which cannot be decoded. Making this division also reduces the cost of writing all data on blockchain. Along with these, the system supports INR payments as well for ease of access.

2 Related Works

The data from the transactions is dispersed and shared throughout a network that connects millions of computers. Blockchain is a decentralised solution that eliminates the need for a third party intermediary. Nodes on the network validate and encrypt this information. In addition, as the network of nodes grows longer, it becomes more difficult to hack [5]. Customers' trust in e-commerce firms will increase as a result of the implementation of blockchain [6, 7].

Cooperative assaults by greedy miners can compromise the blockchain. Some believes nodes with more than 51 per cent computing power can reverse the blockchain and undo the transaction in general. Even nodes with less than 51 per cent capacity, according to the recent research, can be helpful. Defensive mining is still profitable for attackers with less than 25% of the computing resources. Consumer-to-consumer transactions model reduces the central position of multinational corporations in regulating and imposing controls and promotes the growth and globalisation of this form of trade.

All of this information may now be stored in a data block in a cryptographic manner. Both the buyer and the store will benefit from this change [7]. For example, the blockchain customer database may be used to redeem rewards and for warranty and guarantee claims, as well as for promotions and loyalty schemes [8–10]. A high level of authenticity and dependability would be required for the redemption of reward points [4]. The success of the blockchain is being used to entice investors by offering a large return. Until incorporating this technology into a business solution, it is necessary to determine if it meets the specifications. Combining big data analytics with blockchain can solve many problems easily especially in data management and data analytics [11].

Smart contracts allow contracts to be executed automatically depending on predefined criteria and conditions. With its integrated management systems, smart contract technology may help simplify process. Due to its ability to automate essential e-commerce operations according to pre-established standards, it is able to eliminate the need for human resources up to a certain point. A smart contract, for example, may be used to control inventory [12]. As well as preserving inventory records, it allows for automated replenishment. It will result in a zero out of stock situation for online stores, increasing convenience for consumers [6].

Blockchain-based platforms have lower transaction fees than traditional e-commerce platforms. This opens the door to micropayments that can be made in a flash. Aside from being faster than traditional payment systems, blockchain transactions in e-commerce are also cheaper. According to a research by Monetha, the traditional payment method involves 16 steps and charges clients between 2 and 6 per cent in processing costs [8].

Platforms built on blockchain technology, such as Ethereum's lighting network, can process millions of transactions per second at very low transactional costs [8–10]. For e-commerce, this technology has the potential to deliver lower transaction costs, stronger security requirements, and a smoother consumer experience [8].

Blockchain's distributed ledger features, combined with its unbeatable features, make it a very appealing technology for solving existing problems form industries, i.e. from both financial and non-financial industry problems. There are many innovative blockchain-based business applications, and as a result, a large number of start-ups that have been working on them [13]. Blockchain systems are more secure. An attacker would need to hold 51 per cent of the blockchain network's global hash power to compromise it, which is incredibly costly and not long-term, viable. This is how blockchain eliminates fraud and double-spending, letting everybody to act as their own bank without the need for an intermediary.

Considering all the features of blockchain integration to e-commerce applications and the benefits it brings to the table motivated us to work on our proposed implementation. In the current scenario, where people are getting increasingly concerned about how strongly monopolies affect their livelihoods and money. We believe that our proposed solution will create a divergence and people will start believing in the applications of blockchain technologies and their applied use cases.

3 Problem Statement

The problems we aim to solve span over various issues faced by popular e-commerce platforms.

3.1 Security Issue

Hackers are always a threat to centralised e-commerce platforms that keep user and product data on a server or in the cloud.

3.2 Ethical Practice

Centralised platforms possess authoritative powers over the data and there is no transparency as far as their functioning is concerned. Popular e-commerce platform has algorithms that boost branded products in their catalogue overshadowing any other product created by small businesses. Other than this, it becomes very tough for local businesses to follow tough procedures of KYC and verifications that the banks and popular e-commerce websites impose.

The small businesses rely on these merchants, but they do not have much leverage on their own. Sellers claim that if their things are discovered for fewer prices on other sites like Walmart, Amazon essentially punishes them by lowering their search ranks, for example. Small companies, in particular, may believe they have little choice but to comply with any requests if the majority of their sales come from these big established e-commerce platforms.

3.3 Issues in Complete Shift to Blockchain

Shifting e-commerce completely to blockchain also has its issues, hence, there has to be a way to add the features of blockchain without making it very difficult for the user to operate through the website. Added with the cost of gas fees (transaction fees) that every read/write transaction take is large. Hence, it is challenging to implement blockchain with solving the issues associated with it.

4 Blockchain Technology Application in E-Commerce

Digital ledger of Ethereum blockchain records immutable blocks of transactions. This is seen in Fig. 1, where the public key, private key, previous hash (and transaction data) is combined to form a hash for each transaction. The public key serves as an address for participants to identify one another, while the private key is utilised to

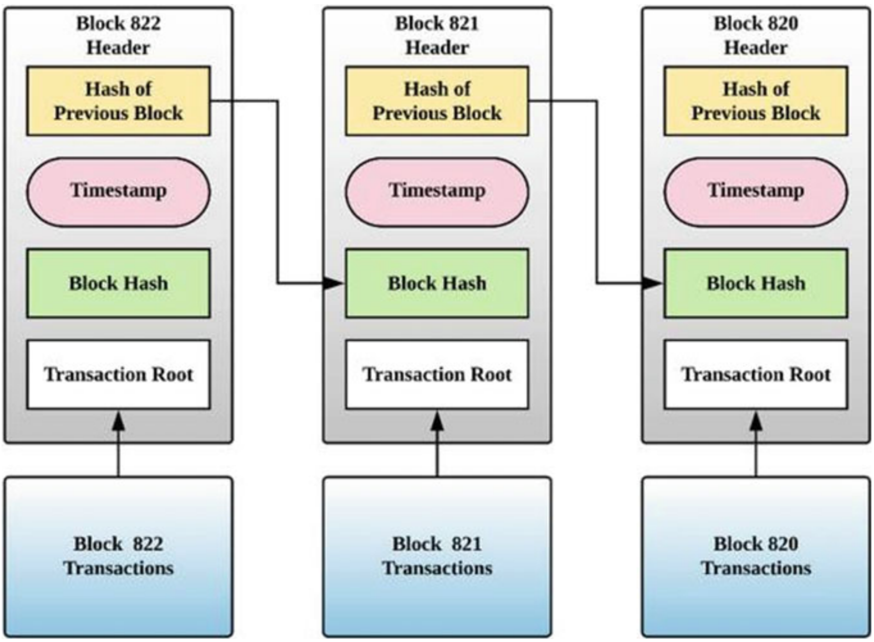


Fig. 1 Blockchain data structure

authenticate them. To connect the blocks, the prior hash is saved. All nodes in the distributed peer-to-peer network are notified of this transaction before it is included in a block. Only once a transaction is approved by a majority of nodes will be added to a block in the mainchain.

There are many terminologies which play a significant role in the field of blockchain applications. The significant technologies and concepts used in the proposed e-commerce site are explained as follows.

4.1 Metamask

Metamask being a cryptocurrency wallet interacts with the Ethereum network through software. It solves the purpose of accessing the Ethereum wallet by utilising a browser extension or a mobile app to, which can then be used to engage with decentralised apps. Metamask allows us keep and manage account keys, broadcast transactions, send, and receive Ethereum-based currencies and tokens, and securely connect to decentralised apps using a web browser or the mobile app's built-in browser.

4.2 EVM and Solidity

The Ethereum Virtual Machine is the Ethereum smart contract runtime environment. The Ethereum Virtual Machine focuses on providing security and allowing machines all around the globe to execute untrusted programmes. The EVM is specialised in preventing DOS assaults and assures that applications do not have access to each other's state, allowing communication to take place without interruption. Solidity is a high-level programming language for constructing smart contracts that is contract-oriented. Solidity was built with the Ethereum Virtual Machine (EVM). Solidity is a statically typed programming language that allows inheritance, libraries, and sophisticated user-defined types.

4.3 Ethereum

Along with its own currency (ETH), Ethereum features its own programming language (solidity). In the form of a blockchain network, Ethereum is a decentralised public ledger for validating and recording transactions. Ether is the network's cryptocurrency. Users of the network may create and publish apps on the platform, monetise them, and use them. Internally, decentralised apps on the network are referred to as "DApps". Ethereum is the second most valuable cryptocurrency, behind Bitcoin, in terms of market capitalisation. According to Ethereum's architecture, smart contracts and distributed apps (DApps) might be created and published without the risk of downtime and fraud, as well as third-party interference. On the Ethereum network, **gas fees** relate to the computational effort required to complete specified actions. The successful completion of a transaction on Ethereum requires the payment of ether (ETH, +4.38%).

4.4 Smart Contact

It is a software programme or transaction protocol that automates the execution of legally significant events and acts in accordance with the provisions of a contract or an agreement, as well as their documentation [14–17]. Reducing the need of intermediaries in processes of arbitration, enforcement costs, fraud losses, and intentional and handling inadvertent exceptions are among the aims of smart contracts [15]. Multiple cryptocurrencies, including Ethereum, now support scripting languages, allowing for the creation of more complex smart contracts between untrusted parties [7] The terms smart contracts and smart legal contracts should be distinct from one another. In the second case, terms are defined and implemented in machine-readable code, of a typical natural language contract.

4.5 Cryptocurrency

A cryptocurrency is a digital asset. To protect transactions and verify coin ownership, a cryptocurrency uses powerful cryptography to keep individual coin ownership information in a ledger. With the rise of cryptocurrencies, a growing array of blockchain technologies has emerged, each with its own set of goals and use cases. Table1 shows the difference of usage between the most used Ethereum and Bitcoin. We have used Ethereum for our proposed model which is being used for 15–25 transactions per second.

5 Proposed System

Our approach is based on the creation of a secure market for everybody. This will provide fantastic opportunity for small companies attempting to compete in a market

Table 1 Comparison between cryptocurrency

Technology	Transaction/sec	Security
Bitcoin	7	Medium
Ethereum	15–25	Medium

dominated by centralised e-commerce platforms. It aims to solve issues of security, transparency of functioning, and ease of money flow.

5.1 Features of Our System

a. The Fast KYC system for registration of users is one benefit of the proposed system. This type of KYC registration is completed in simple and easy steps and can be done from any convenient place.
b. A user can buy or sell as per requirement and wish.
c. A smart contract is a self-enforcing contract that is controlled by a blockchain. Using smart contract in our system reduces the formality and expense associated with traditional approaches while maintaining authenticity and credibility. The code consists of a set of rules that parties in a smart contract agree to follow. So long as certain conditions are met, the agreement will immediately take effect. These regulations are known to the public, which reduces the likelihood of unfair actions and corruption occurring. Hashing is used to connect blocks, which are used to keep track of transactions.
d. Though, on the blockchain, privacy and the cost of gas, which is necessary to write any data on the blockchain, become issues, our proposed approach is both quick and secure. We have merged MongoDB with blockchain for data storage, which makes our solution unique. This strategy distinguishes our system from others. It not only minimises the gas cost of writing data on blockchain, but it also secures the system since data split on both. We keep just the Name and UID, as well as certain critical information, on the blockchain, and the rest of the data on MongoDB, with a Metamask ID hash code connecting the two. As a result, when the attacker breaks into the server, he will only have access to irrelevant data.

6 System Design

Our system is completely transparent and secure system. The easy flow diagram of our proposed system is depicted in Fig. 2 with the feature mentioned before. A customer can buy and sell on its discretion.

6.1 Implemented Modules

The proposed system can be implemented in separate modules. The following is the module division for our system.

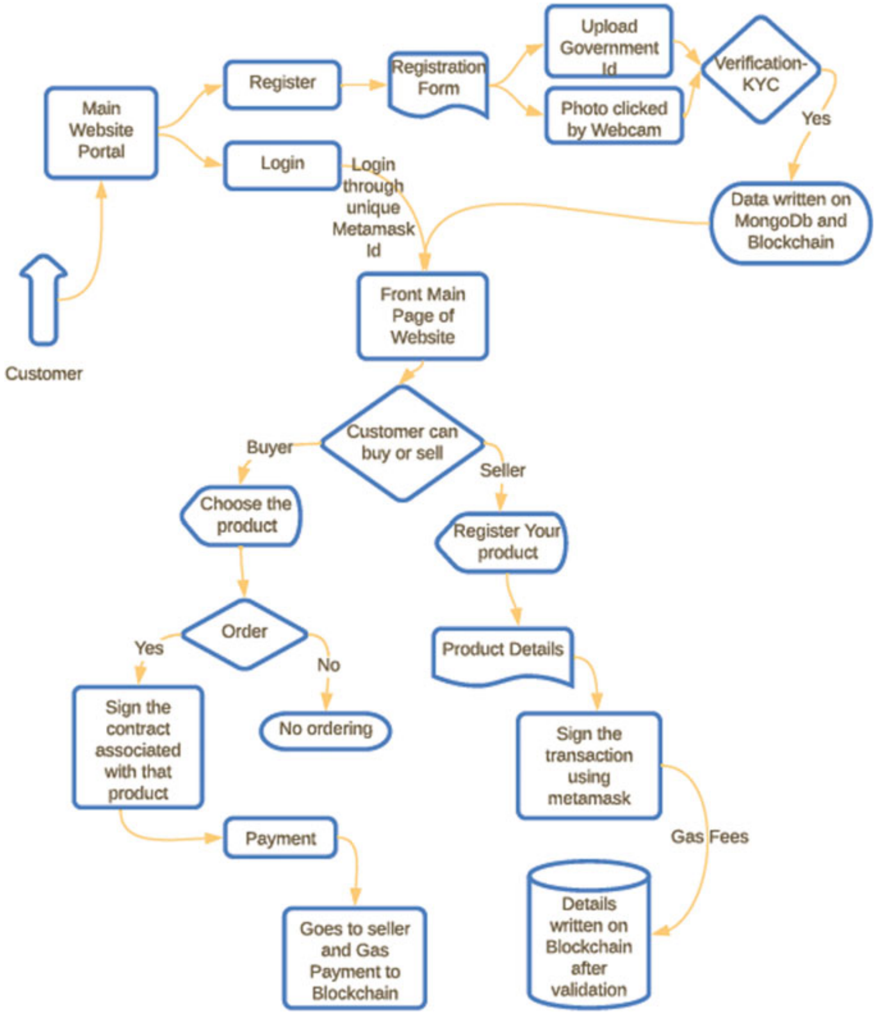


Fig. 2 Flowchart of proposed system

6.1.1 Fast KYC (Know Your Customer)

Each organisation must verify user identification in some way, and this is especially vital for financial institutions. KYC procedures arose as a result of this to aid businesses in ensuring they know who they are doing business with. Typically, this is a lengthy, drawn-out procedure in which certain user information forms are displayed and various types of background checks or verification are performed.

Verification is a vital and initial stage in our system; hence, we have included a Fast KYC method. Figure 3 depicts the complete procedure of verifying user Ids before starting any process on the blockchain model. Initially, images are captured

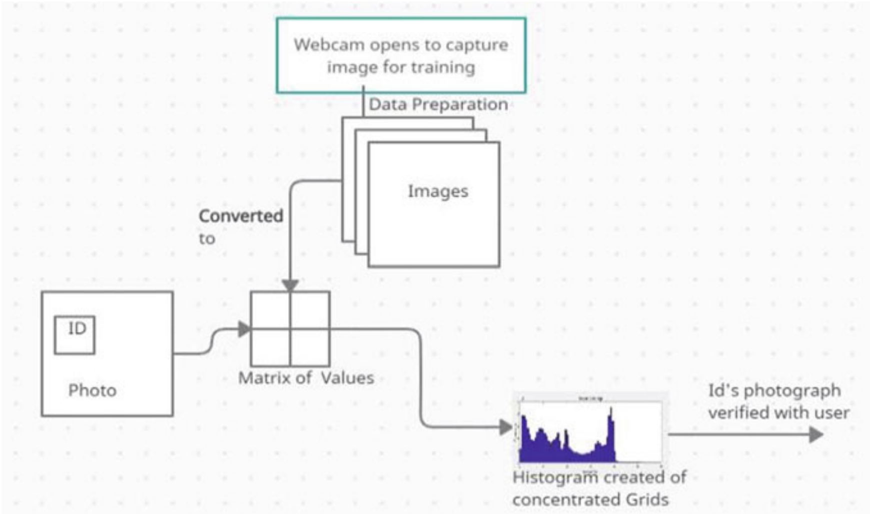


Fig. 3 Fast KYC system user registration

through normal webcams which are simultaneously processed and trained using stream processing.

Stream processing being a type of real-time processing helps to perform operations on data as it is being generated. The multiple trained images are tested upon the user's Id photo and using Haar classifier. Haar Cascade classifiers are an effective way for object detection. As a result, Id photographs are compared to webcam dataset pictures using a histogram constructed from the dataset's concentrated grids. Once verified, Id is parsed using tesseract to find text and then returns Id number for administer verification. This module enables vendors to register on an e-commerce platform with the utmost convenience and in the shortest time possible. Sellers can register their products on the site when the verification procedure is done.

6.1.2 Login/Registration

Metamask (Ethereum wallet) plays a vital role throughout this module. A user must sign in to their account only using their own wallet in addition to providing their details such as email and password. There are restrictions for possessing multiple accounts for single user. A public–private key pair is generated when an owned Ethereum account is created. The account's address is given by the final 20 bytes of the public key. This address is used for token transfers and account association. To govern the account and create transactions, a private key is used. It is always encrypted and kept in a keystore. To minimise risk of data leaks, we are dividing the user data in two parts and they will be stored in MongoDB and the mainchain in parts. This will help as a hacker to get hold of important data of the users have to successfully crack and decode data from both data storages as well as join them. This feature will make the data even more secure. To crack or make unauthorised access to any user's account too, the hacker will not only require the Email and Password, but also need to get hold of the Users Wallet address private key and seed phrase. The login and registration part implement the unique features making the proposed system better than existing one. The user-friendly interface makes the design more usable.

6.1.3 Linking Exchanges Through Binance API

An API key may be viewed of as a username that is generated to facilitate data access. The exchange generates the Binance API in our system, by which user can then feed on to another application. Based on the permissions user provides, the programme will import your data. It provides real-time Ethereum prices and also a linkup to the cryptocurrency exchange portal, so while using our application, user can add ETH into his wallet or get fiat tokens from the exchange itself. After each transaction or investment, the portfolio will begin to update automatically. It will help the users to view the analytics report, check their transaction and account details, create exchange-specific notifications, and do other things all in one place.

6.1.4 Listing and Marking Products

Both data write modules are completely executed on the mainchain providing maximum transparency to the fact that each product listed on the Platform is provided with equal benefits and there is no company interference in it. Due to the fact that data is stored in a public database, and it cannot be amended, and the price amount may be transferred automatically without the participation of a third party, thanks to the blockchain. As a result, smart contracts ensure that pricing or regulations that have been set and signed do not unexpectedly alter. Digital rule agreement between two parties is called a smart contract.

Figure 4. shows complete procedure for product registration on seller side. For product registration, required details are filled in the registration form like any normal process which is accessed and processed through Metamask. Once the transaction is signed by the seller, storage-write transaction requests are accepted by Ethereum chain using smart contracts. These contracts are fixed and written on the blockchain simultaneously, gas fees is paid to the validator after complete verification and write process.

6.1.5 Buying and Making Payments

Ether is a coin that is used by the Ethereum network to pay for gas. When a new valid block is introduced to the network, ether is generated and rewarded to the node that

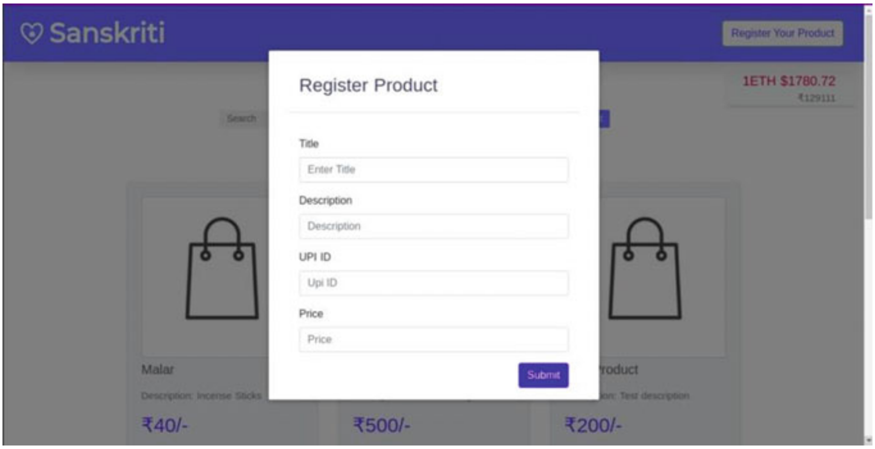


Fig. 4 Product registration by seller

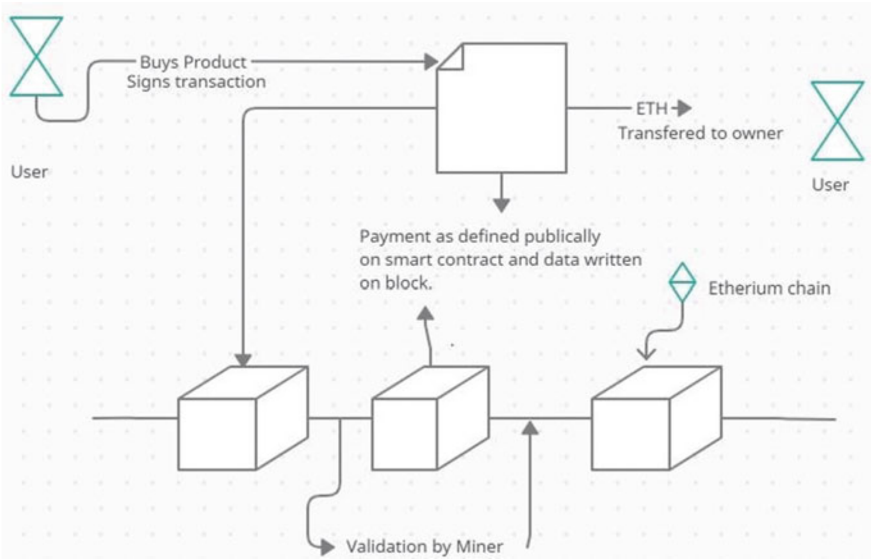


Fig. 5 Buy product use case

created the block. Gas is a computational unit that is utilised for all Ethereum state transitions such as transactions and computational services. When a transaction is made, it is signed using the owner's private key. The private key is required throughout the validation procedure to validate that the sender is the owner of the account. This reduces the chances of frauds and discrepancy.

Figure 5 depicts the ledger system which helps us to create a consumer–platform–consumer relationship.

Once the buyer has logged in, he or she can order any of the offered/listed items with a single click. Following the buyer's signature on the purchase transaction, the product's contracts are examined. Following that, payments are specified openly on a smart contract in order to present payment-related information to the buyer, and data is written to the chain at the same time. Once the transaction has been validated and signed by the blockchain miners, the user and seller are connected directly. The cost is paid by the buyers, and the Ethereum money equivalent is sent to the direct seller according to the smart contact rule.

7 System Requirements

The project can be easily carried out on any low-end PC with Operating System Windows 7 or above version 64-bit CPU Intel Core i3 having RAM of minimum of 2 GB. The project requires installation of Metamask extension along with Ganache from Truffle Suite which is Local Ethereum chain. A proper connection of Metamask to the local Ganache Ethereum chain is established.

8 Results and Discussion

We have successfully implemented the proposed model which allows local sellers to register easily with valid government Id. The transaction is secure process with full transparency using Ethereum cryptocurrency and Metamask. The results are quite positive which can be useful for the society.

The main portal for the proposed e-commerce provides option of Register to new users and Login to existing users. Figure 6 shows the first main page with both the options.

For the existing users, login is verified against the stored Metamask Id and password. Login is allowed only after connecting to relevant Metamask account. For the register case, new Ids are created after Id verification. All the details except name and Metamask Id are stored on the MongoDB database as shown in Fig. 7. User database contains all the main User data but without any names, the names are written to the blockchain along with Metamask wallet Id that links both the data.

Ethereum wallet is in itself very secure in terms of usage, has a 12-worded seed phrase that a user needs to remember as a password, making it very secure. Figure 8 shows the unique Metamask account of a customer with available ETH value.

Our proposed work uses this feature as a benefit by linking the wallet Id to the login security. Hence, a hacker not only needs to a UID and password to access any account, but also needs to crack or get hold of the wallet to enter our application.

Figure 9 shows the homepage which maintains the list of registered products. This model serves as customer-to-customer application.

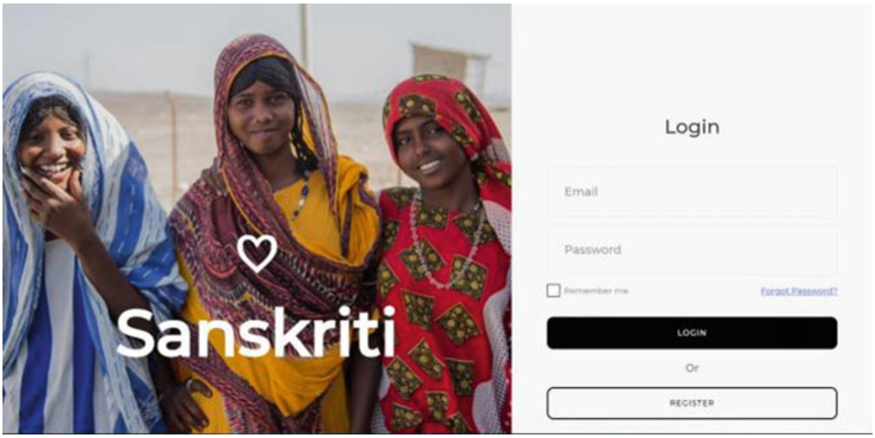


Fig. 6 Main page of the website

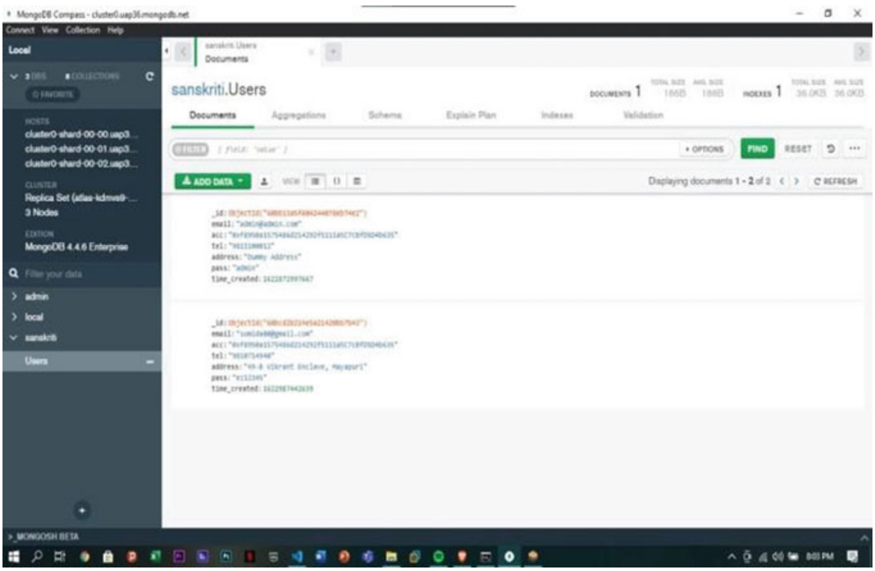


Fig. 7 Mongo DB database storing User's details

Any user can opt to sell or buy product depending on requirement. For selling purpose, one needs to click the "Register Your Product" on the right in Fig. 9. Figure 10 shows the form accepting the product details. All entries within the application go through the decentralised procedures. The framework provided by Ethereum helps us define publicly visible rules (smart contracts). Hence, whenever any action takes place such as this (listing of product) the contract needs to be signed and the gas (validator's fee). It is ensured that the process is simple and less time consuming.

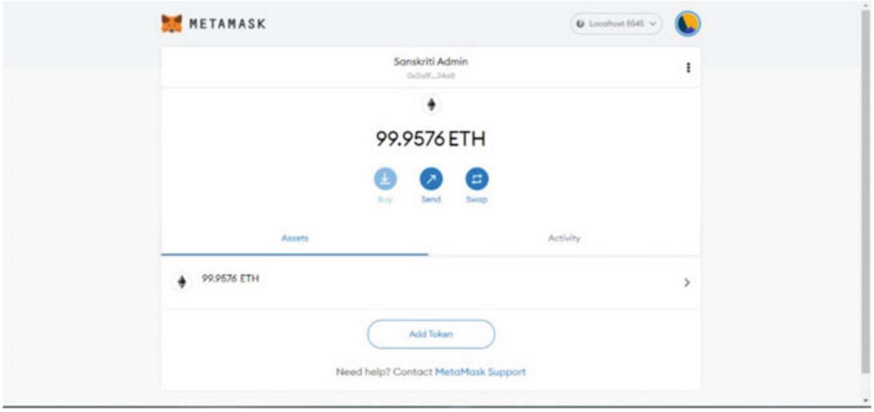


Fig. 8 Metamask (Ethereum wallet)

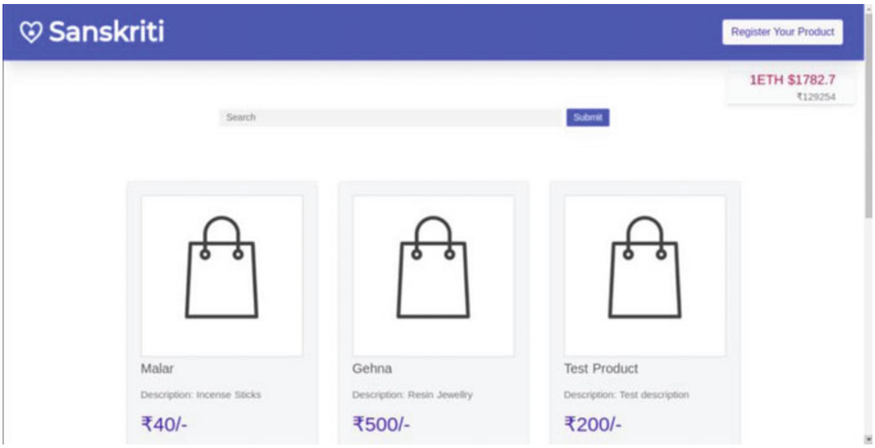


Fig. 9 Page showing list of registered products

For the case of buy order too, the contract need to get e-signed the same way, just this time, as it is specified in the contract that the amount needs to be transferred hence the wallet opens up with the payable amount prefilled in it, on signing the amount along with the details of user (only parts required for completing the order) who wanted to buy will be written on the chain for transparency, and the said amount will be transferred. Figure 11 shows the user interface for ordering the product for any buyer. The interface is made user-friendly to reduce complications with increased reliability and security.

We have successfully built a platform which is more secure due its unique feature of distributing data on MongoDB and the main blockchain. The gas cost is comparatively less for our system as the content written on block is reduced due to us- age

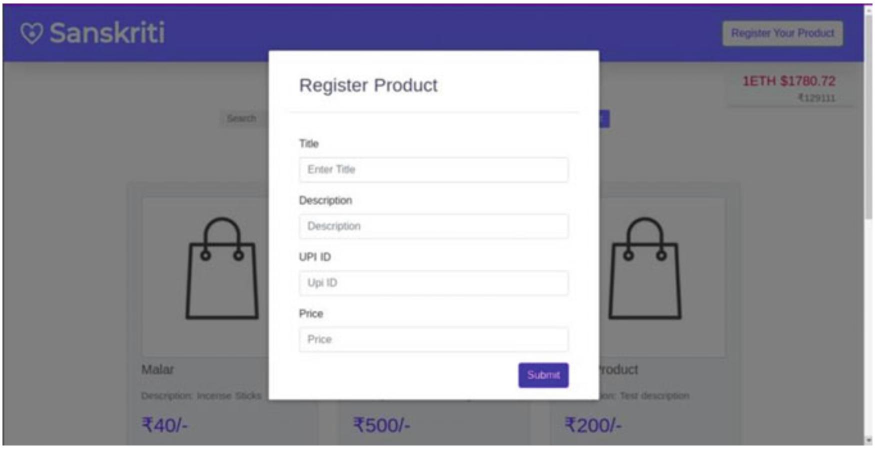


Fig. 10 Registration of the product

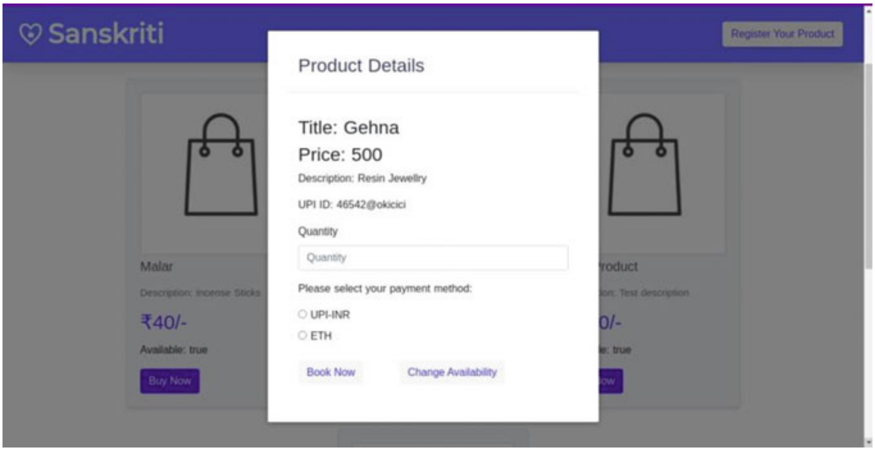


Fig. 11 Product detail viewing and ordering

of MongoDB Atlas database. These mentioned features solve the main issues faced by the existing e-commerce sites.

9 Conclusion and Future Work

Blockchain e-commerce use cases including cryptocurrency payments, inventory, and data security demonstrate how technology benefits both customers and sellers. It enables reduced pricing to be established, effective asset monitoring, privacy to be protected, and the avoidance of unnecessary expenses. This blockchain technology is bringing huge change in the economics globally. This new concept is much better than the traditional systems in many aspects. Our research work suggests solution to the problems faced by the existing e-commerce sites. This work can contribute in making an efficient and cheaper system for online selling platforms. Our proposed system is fully implemented model which consumes less time and provides benefit of Fast KYC registration and better security policies. Blockchain refers to a growing collection of records known as blocks that are linked to the usage of encryption. Each block includes prior block encryption segmentation, a time stamp, and transaction data. Through this project, we aim for promoting business of local people. The significant transparent feature makes this more reliable and useful.

This research can contribute significantly to the literature of advanced technology and e-commerce in blockchain. Our future work will be including extending the support for various other crypto currencies and making a platform specific token, so that people who do not understand crypto can deposit into the application and all payments will be done via the platform token. This will improve the model more significantly.

References

1. An introduction to blockchain in ecommerce. https://inviqa.com/blog/introduction-blockchain-ecommerce
2. Send DC, Cox D, Has the blockchain technology made e-commerce business easier?. https://www.cryptonewsz.com/how-can-blockchain-transform-e-commerce-and-the-way-of-doing-business/
3. How blockchain has helped the business of e-commerce. Asia Blockchain Review—Gateway to Blockchain in Asia. https://www.asiablockchainreview.com/how-blockchain-has-helped-the-business-of-e-commerce/
4. Blockchain and e-commerce: How do they make a winning .... https://www.orangemantra.com/blog/blockchain-e-commerce-make-winning-combination/. Accessed: 23 March 2020
5. Geer D, How blockchain will transform the e-commerce industry over the next few years. https://thenextweb.com/contributors/2018/06/15/how-blockchain-will-transform-the-e-commerce-industry-over-the-next-few-years
6. Zheng W, Zheng Z, Chen X, Dai K, Li P, Chen R (2019) NutBaaS: A Blockchain-as-a-service platform. IEEE Access 7:134422–134433. https://doi.org/10.1109/ACCESS.2019.2941905
7. Sharma TK, Top 10 blockchain solutions for e-commerce. https://www.blockchain-council.org/blockchain/top-10-blockchain-solutions-for-e-commerce
8. Can blockchain change e-Commerce in 2020? https://marketrealist.com/2019/12/can-blockchain-change-e-commerce-in-2020/
9. Post C, Blockchain technology to revolutionise e-commerce businesses in Australia. Business Partner Magazine. https://businesspartnermagazine.com/blockchain-technology-revolutionise-e-commerce-businesses-australia/
10. Schelén MO, Andersson K (2019) A survey of blockchain from the perspectives of applications, challenges, and opportunities. IEEE Access 7:117134–117151. https://doi.org/10.1109/ACCESS.2019.2936094
11. How blockchain is impacting ecommerce. Ignite Ltd., 08 Feb 2019. https://igniteoutsourcing.com/blockchain/how-blockchain-is-impacting-ecommerce/

12. Zheng Z, Xie S, Dai HN, Chen X, Wang H (2018) Blockchain chal- lenges and opportunities: A survey. Int J Web and Grid Services 14(4):352–375
13. Röscheisen M, Baldonado M, Chang K, Gravano L, Ketchpel S, Paepcke A (1998) "The Stanford InfoBus and its service layers: Augmenting the internet with higher-level information management proto- cols". Digital Libraries in Computer Science: The MeDoc Approach. Lecture Notes in Computer Science. Springer. 1392:213–230. https://doi.org/10.1007/bfb0052526.ISBN978-3-540-64493-4
14. Fries M, Paal BP (2019) Smart contracts. Mohr Siebeck
15. Savelyev A (2017) Contract law 2.0: 'Smart' contracts as the beginning of the end of classic contract law. Info Comm Tech Law 26(2):116–134. https://doi.org/10.1080/13600834.2017.1301036
16. Tapscott D, Tapscott A, (2016) The blockchain revolution: How the technology behind bitcoin is changing money, business, and the world. pp 72, 83, 101, 127. ISBN 978-0670069972
17. Szabo N (1997) View of formalizing and securing relationships on public networks | First Monday. firstmonday.org

Classification of Animals Using MobileNet with SVM Classifier

M. Sowmya, M. Balasubramanian, and K. Vaidehi

Abstract This paper proposes classification of different animal images comprising dog, cat, elephant, panda, and monkey. The most significant task in computer vision and deep learning is animal image classification. This work helps the forest service in further developing their alarm systems to the people who stay close to the forest area. Animal classification is beneficial in research allied to focused animal behavior, and also dangerous animal disturbance in domestic location can be avoided. This paper is focused on MobileNet Architecture which is implemented over Animal Classification Dataset for categorization of distinct animals. The extracted features from MobileNet Architecture are then used to classify the different categories of animal images with the utilization of an SVM classifier. Extensive experimentation on an Animal Image Dataset of 800 images with various classes are used to evaluate the performance. From the results of MobileNet with SVM has achieved a good performance of 99%.

Keywords Animal image classification · Deep learning · CNN · MobileNet · SVM

1 Introduction

Animal image classification is a significant task for signifying the presence of animals on highways or in residential areas. Accidents on highways near forest area are primarily caused by wildlife animals on the road. When these activities occur at night, traveling across these roads becomes extremely difficult, and it becomes a huge obstacle for travelers as well as surrounding communities or human settlements. The challenge of identifying and classifying animals is quite difficult, and thus, it leads to animal classification. The biologist had to manually classify the animals in order to observe their behavior which was a difficult and time-consuming

M. Sowmya (✉) · M. Balasubramanian
Department of CSE, Annamalai University, Chidambaram, India
e-mail: msoumya@stanley.edu.in

K. Vaidehi
Department of CSE, Stanley College of Engineering and Technology For Women, Hyderabad, India

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_25

task. The classification and detection of animals is a challenging research topic in computer vision. The difficulty with the automatic approach is that animals do not show separately at a scene; instead, all of the natural objects in the area are taken into consideration. As a result, animal image segmentation, animal image classification, and animal behavior detection are the open research areas in computer vision.

Animal classification has various applications, including animal-vehicle accident prevention, animal track facility, identification, antitheft, and zoo animal security. Camera traps are common in ecological and management studies for remotely viewing wildlife and are considered one of the important and powerful tools for wildlife study. Camera trap studies generate millions of images that must be examined, often by visually inspecting each image, in order to extract data for ecological analyses.

Animal classification is important in different professions and activities, including farming, railway track monitoring, highway monitoring, and so on. Constantly monitoring and classifying an animal by spending hours is an extremely difficult task. The problems with manual classification and detection of animals include the need for a human intervention and humans have to constantly watch the region 24 h a day, which is not feasible and would also be prone to human error. For this form of monitoring, an automated image classification is required because it would make the process easier and reduce human work. CNN have had enormous success in image classification and object detection. Support Vector Machine (SVM) is one of the most important Machine learning approaches which can be utilized for categorization and regression. In this work, describes the MobileNet CNN architecture for extracting features and SVM for classifying the distinct images of animals. MobileNet is a CNN architecture that is both efficient and portable, and it is employed in real-world applications.

2 Related Work

In this literature, distinct methods for automatic animal image classification have been proposed. The work in [1] uses the method, consists of two sections, and uses mixture of methodologies (SIFT—Scale-invariant feature transform, SURF—Speeded Up Robust Features). On a few images from the dataset, the usefulness of the provided hybrid approaches is illustrated. The wolf, fox, brown bear, deer, and wild boar are among the major animals found on Slovak country. The method given here could be utilized to other fields of image classification and feature extraction. The results of the experiments suggest that using the arrangement of local descriptors improves object detection. The paper [2] concentrates on advances in graphics processing units (GPUs), parallel computing, and deep neural networks which have resulted in major development in the fields of machine learning and computer vision. The objective of this paper is to study the use of a convolutional neural network to classify animals. The convolutional neural network is a strong machine learning technology that is trained on a huge number of different images. In this paper, a convolutional neural network and a support vector machine are used to classify animals. AlexNet-pretrained convolution neural network is used to train the animal images. The features which are extracted are then sent into the multiclass SVM classifier for classification. To estimate the performance, the authors have conducted wide experimentation on their own dataset which contains 500 images with 50 classes and each class contains 100 images. The proposed method which the authors used has achieved good classification rate [3]. The authors have developed a supervised, based classification system for the classification of real-world animal images. The classification model is trained using toy images of animals. The image is preprocessed to eliminate noise before being enhanced with adaptive histogram equalization and median filtering. The preprocessed toy image was then segmented using the k-means clustering algorithm in the second stage. The toy animal image is separated from the background using segmentation. The extraction of hog features from the segmented image is the third stage. The extracted features are then put together in the final stage.

The description of the paper [4] is about the Dogs versus Cats contest on Kaggle is attempting to overcome the CAPTCHA problem, which involves discriminating between dog and cat photos. It is simple for humans, but evidence suggests that automatic cat and dog separation is very challenging for cats and dogs. To address this problem, several people have been working on developing machine learning classifiers. On the Asirra dataset, a color-based classifier had an accuracy of 56.9%. Based on color and texture features, an SVM classifier attained an accuracy of 82.7%. They used SIFT (Scale-Invariant Feature Transform) features to train a classifier, and the result was 92.9% accuracy. To address this issue, the authors had tried a few different approaches. They experimented with Dense-SIFT features, merging Dense-SIFT and color features, and CNN features, for example. They also employed SVMs on the learned features and were able to attain a classification rate of 94.00% [5]. Authors described system for classifying animal behaviors from fixed-point camera recordings from Videos. It is required to detect and monitor animals in order to classify their behavior. Background subtraction and frame subtraction are two methods for recognizing moving objects. Because they are subject to sunlight and shadow, approaches are ineffective for detecting animals kept indoors. Using information provided by DeepLabCut, authors described a method for tracking animals and classifying their behavior. The proposed method outperforms the standard method, according to the results of the experiments.

The work in [6] Examined Digital image recognition has been utilized in a variety of applications; the most common of which are object classification and detection. Image recognition monitoring of animal activity in natural areas is critical for animal health and production. Sheep Pain Facial Expression Scale (SPFES) is now being used to monitor sheep's facial expression. The models were prepared using data augmentation, L2 regularization, and fine-tuning. Using the VGG-16 model, the experimental findings for the sheep facial expression dataset achieved 100% training, 99.69% validation, and 100% testing accuracy. The paper [7] is focused on CNN and present results obtained in training a CNN to classify 20 African wildlife species with an overall accuracy of 87.5% from a dataset containing 111,467 images. The authors in paper [8] described about feature extraction which is a critical stage in the detection of animal fiber microscopic images in the wool and textile industries. A hybrid model based on Convolutional Neural Network (CNN) and Random Forest (RF) is proposed for automatic feature extraction and classification of animal fiber microscopic pictures to improve the accuracy of wool and cashmere microscopic images classification. The work described in [9] using deep learning to examine body language of household animals, and this paper determines whether they are unwell or not and provide timely assistance with animal species classification in order to achieve this goal. This paper is focused on VGG, a deep learning model, to distinguish between cat and dog images. After utilizing VGG-16's Transfer Learning, the accuracy has been increased from 80 to 95%. The paper concentrates [10] on extremely difficult camera trap image data, and this paper is focused on deep convolutional neural network-based species recognition technique for wild animal classification. The imagery was taken using a motion-triggered camera trap and automatically segmented using a cutting-edge graph-cut algorithm. 14, 346 training images and 9, 530 testing images in the typical camera trap dataset of 20 common North American species are used.

The journal [11] described about goat breed database was constructed using goat images from six distinct breeds that were collected from several organized registered goat' farms in India. Nearly two thousand digital images of goats were taken without causing stress to the animals. Inception-v3 has been found to perform better than VGG-16 in terms of accuracy and training duration. The paper [12] concentrates on Digital imagery from a camera which can be used to study marine species' behavior, seasonal distribution, and abundance. The animals are classified as characteristics (shape, color, texture, etc.) or dissimilarity matrices created using various shape analysis algorithms, depending on the quantity and attributes of acquired images. This paper is projected on the cost function for multiview learning like correntropy loss, which has good statistical features for rejecting noise. The work in [13] describes about elephant photo identification method based on the shape comparison of the nicks that characterize elephant ears presented in this research. The technique used in this work can deal with photos that are congested and noisy, such as those used by zoologists to identify wild elephants. The paper [14] describes that from a few training images, the approach dynamically builds a color model of elephants. Location of elephants in video sequences with varying background and lighting conditions using the color model is used in this paper. Elephants (and groups of elephants) of various sizes and attitudes conducting various activities are detected using the proposed method.

The paper [15] projects about how to categorize the 1.2 million high-resolution photos in the ImageNet LSVRC-2010 contest into 1000 separate classes, and a deep convolutional neural network was used. They achieved top-1 and top-5 error rates of 37.5% and 17.0%, respectively, on the test data, which is better than the prior state of the art. The work in [16] is focused on recognizing degraded character images recovered from ancient Kannada poetry sources, as well as handwritten character images collected from a variety of unconstrained situations. The training of degraded patterns of character picture samples is done out using AlexNet, a deep convolutional neural network. Handwritten datasets obtained synthetically from users in the age categories of 18–21, 22–25, and 26–30, as well as printed datasets recovered from ancient document pictures of Kannada poetry/literature are used to evaluate the performance of this project.

According to [17], the work describes how the accumulation of solid trash in metropolitan areas is becoming a major worry, since it will pollute the environment and pose a threat to human health. The extractor is created using the 50-layer residual net pre-train (ResNet-50) Convolutional Neural Network model, which is a machine learning technique, and the Support Vector Machine (SVM), which is used to categorize garbage into distinct groups/types such as glass, metal, and plastic. The proposed system is tested on the garbage picture dataset created by Gary Thung and Mindy Yang, and it achieves an accuracy of 87% on the dataset. The paper [18] concentrates on Hybrid convolution neural network utilized with transfer learning (TL) and support vector machine (SVM) to recognize four types of wear debris, including cutting, spherical, fatigue, and severe sliding particles, for the aim of automatic wear debris categorization. The proposed hybrid CNN with TL and SVM outperforms the CNN, CNN with TL, and CNN with SVM in terms of classification accuracy and efficiency. The authors in [19] examined about the fetal heart rate (FHR) which is crucial in determining the health of the fetus. Computer technology is essential for the study of FHR in electronic fetal monitoring due to the rapid growth of computer information technology (EFM). FHR is classified into three categories: normal, suspect, and abnormal. They also conducted a relative experiment in which the FHR features were extracted using a feature extraction method based on statistics. The features were then used as input to categorize using support vector machines (SVM) and multilayer perceptrons (MLP). According to the experiment's findings, the accuracy of classification of SVM, MLP, and CNN is 79.66%, 85.98%, and 93.24%, respectively.

3 Methodology

This paper concentrates on MobileNet architecture with SVM, and block diagram is shown in Fig. 1. MobileNet model is used as a feature extraction technique, and extracted features are given to SVM classifier to classify images such as dog, cat, elephant, panda, and monkey.

3.1 MobileNet Architecture

For computer vision applications, MobileNet deep Convolutional Neural Networks is performing well. Image categorization is the most common application of MobileNet CNN. This improves accurate CNN prediction efficiency. MobileNet is a CNN architecture that is both efficient and portable, and it is used in real-world applications. In other CNN architectures, feature map is created using all channels of a kernel.

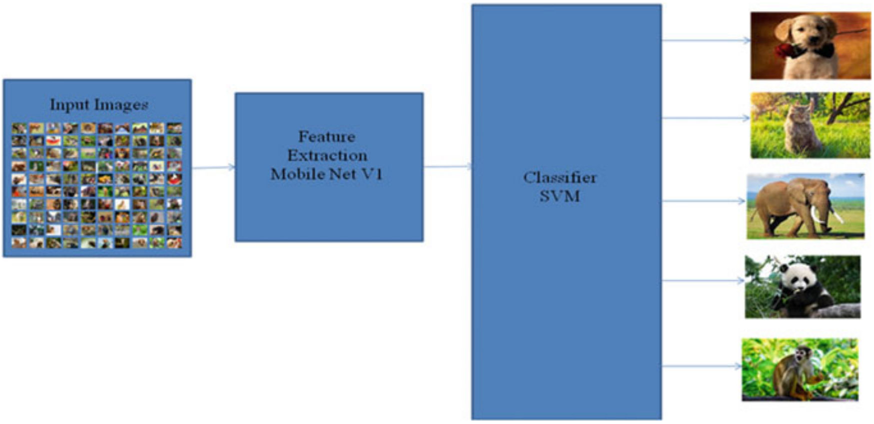


Fig. 1 Block diagram of automatic animal classification

Each channel of a kernel is utilized to build a feature map in a depthwise convolution of MobileNet CNN. MobileNet adds two new global hyper parameters (width multiplier and resolution multiplier) which helps the model developers to trade off latency or accuracy for speed, depending on their requirements. Andrew G. Howard and colleagues created the MobileNet v1 paradigm. MobileNet v1 which can be used to detect and classify images and objects [20]. The depthwise separable convolution architecture was used to create MobileNet. There are two types of separable convolutions in MobileNet: depthwise convolution and 1×1 pointwise convolution [21]. The batch normalization and ReLU are applied to each layer of MobileNet v1 [22]. A depthwise convolution and a pointwise convolution are used to construct a MobileNet convolution layer. A MobileNet architecture contains 28 layers, and depthwise and pointwise convolutions are counted separately. A basic MobileNet has 4.2 million parameters, which can be decreased even further by appropriately setting the width multiplier hyper parameter. MobileNet CNN architecture with input shape $224 \times 224 \times 3$ is applied as an input image. Down sampling has done, and it will reduce into $112 \times 112 \times 32$ where 32 is depth of the filter; it is applied to depthwise first convolution. Pointwise convolution 1×1 with $112 \times 112 \times 64$ is applied to pointwise first convolution. Again, down sampling is used; it will decrease to $56 \times 56 \times 64$ applied to convolution depthwise second convolution. Pointwise convolution 1×1 with $56 \times 56 \times 128$ is applied to pointwise second convolution. This procedure is followed up to last layer that is 13^{th} layer of depthwise convolution and 1×1 pointwise convolution. Global average pooling is used in MobileNet architecture. Total parameters are 4,253,864; out of which, trainable parameters 4,231,976 and non-trainable parameters 21,888 are used in MobileNet architecture (Fig. 2).

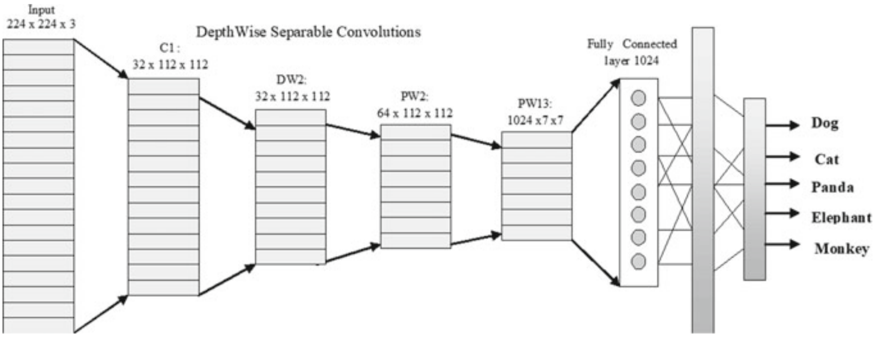


Fig. 2 MobileNet CNN architecture

3.2 SVM

Support Vector Machine is a supervised machine learning technique that can be applied to classification and regression problems. The SVM algorithm's purpose is to find the optimal line or decision boundary for dividing n-dimensional space into classes. Face detection, image classification, text categorization, and other tasks can be used from SVM technique. The SVM is a machine learning technique that was first described in the 1990s and is mostly used for pattern recognition. Many pattern classification issues, including image recognition, speech recognition, text categorization, face identification, and faulty card detection, have been solved using this method. The SVM is typically used for pattern classification, which means that this technique is mostly utilized to categorize various sorts of patterns. There are two types of patterns namely linear and non-linear. Linear patterns are patterns that are easily identifiable or can be quickly divided in low dimensions, whereas non-linear patterns are patterns that are difficult or cannot be simply separated, and so must be further altered to be easily separated. Linear SVM: Linear SVM is a classifier that is utilized for linearly separable data, which implies that if a dataset can be classified into two classes using a single straight line, it is called linearly separable data, and the classifier is named Linear SVM. Non-linear SVM: Non-linear SVM is used for non-linearly separated data, which implies that if a dataset cannot be classified using a straight line, it is non-linear data, and the classifier used is called non-linear SVM. The different types of kernel functions used in SVM are linear, polynomial, radial basis function, and sigmoid.

4 Dataset

Animal Image Classification Dataset contains 12 animal classes like cat, cow, dog, elephant, monkey, panda, sheep, spider, squirrel, horse, hen, and butterfly and 1452 files are there per each category [23]. In this work, 444 images of cat, 634 images of dog, 267 images of elephant, 700 images of monkey, and 699 images of panda have taken for 5 different class animal classification.

5 Experimental Results

In this paper, MobileNet on Animal Image Dataset are used for the better classification performance of the network. From Animal Image dataset, 2744 images from 5 different categories are used; out of which, 80% of the images from these are used for training, and 20% are used for testing the model (Table 1).

In MobileNet CNN architecture, there are 864 parameters in first convolution. First depthwise convolution contains 288 parameters, and first pointwise convolution contains 2048 parameters. Second depthwise convolution contains 576 parameters, and second pointwise convolution contains 8192 parameters. Third depthwise convolution contains 1152 parameters, and third pointwise convolution contains 16,384 parameters. Fourth depthwise convolution contains 1152 parameters, and fourth pointwise convolution contains 32,768 parameters. Fifth depthwise convolution contains 2304 parameters, and fifth pointwise convolution contains 65,536 parameters. Sixth depthwise convolution contains 2304 parameters, and sixth pointwise convolution contains 131,072 parameters. Seventh depthwise convolution contains 4608 parameters, and seventh pointwise convolution contains 262,144 parameters. Eigth, Nineth, Tenth, Eleventh depthwise convolution contains 4608 parameters, and Eigth, Nineth, Tenth, Eleventh pointwise convolution contains 262,144 parameters. Twelfth depthwise convolution contains 4608 parameters, and twelfth pointwise convolution contains 524,288 parameters. Thirteenth depthwise convolution contains 9216 parameters, and thirteenth pointwise convolution contains 1,048,576 parameters. Total parameters are 4,253,864; out of which, trainable parameters 4,231,976 and non-trainable parameters 21,888 are used in MobileNet architecture.

Confusion Matrix of SVM is shown in Table 2. Classification performance using MobileNet with SVM is shown in Table 3. MobileNet is used as a feature extracted in this paper and has achieved 99% accuracy. In this study, the evaluation metrics which are used to evaluate the performance of animal classification such as classification accuracy; Precision and Recall which is indicated in (1), (2), and (3) where P is the total of positive cases and N is the total of negative cases. TP is the true positive quantity. FP is the false positive, FN is false negatives, and TN stands for true negatives.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{P + N} \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

Table 1 MobileNet architecture parameters

Layer (type)	Output shape	Parameters
Input_1	(None,224,224,3)	0
conv1(conv2D)	(None,112,112,32)	864
conv_dw_1(Depth Wise Conv2D)	(None,112,112,32)	288
conv_pw_1(conv2D)	(None,112,112,64)	2048
conv_dw_2(Depth Wise Conv2D)	(None,56,56,64)]	576
conv_pw_2(conv2D)	(None,56,56,128)	8192
conv_dw_3(Depth Wise Conv2D)	None,56,56,128)	1152
conv_pw_3(conv2D)	(None,56,56,128)	16,384
conv_dw_4(Depth Wise Conv2D)	(None,28,28,128)	1152
conv_pw_4(conv2D)	(None,28,28,256)	32,768
conv_dw_5(Depth Wise Conv2D)	(None,28,28,256)	2304
conv_pw_5(conv2D)	(None,28,28,256)	65,536
conv_dw_6(Depth Wise Conv2D)	(None,14,14,256)	2304
conv_pw_6(conv2D)	(None,14,14,512)	131,072
conv_dw_7(Depth Wise Conv2D)	(None,14,14,512)	4608
conv_pw_7(conv2D)	(None,14,14,512)	262,144
conv_dw_8(Depth Wise Conv2D)	(None,14,14,512)	4608
conv_pw_8(conv2D)	(None,14,14,512)	262,144
conv_dw_9(Depth Wise Conv2D)	(None,14,14,512)	4608
conv_pw_9(conv2D)	(None,14,14,512)	262,144
conv_dw_10(Depth Wise Conv2D)	(None,14,14,512)	4608
conv_pw_10 (conv2D)	(None,14,14,512)	262,144
conv_dw_11(Depth Wise Conv2D)	(None,14,14,512)	4608
conv_pw_11 (conv2D)	(None,14,14,512)	262,144
conv_dw_12(Depth Wise Conv2D)	(None,7,7,512)	4608
conv_pw_12 (conv2D)	(None,7,7,1024)	524,288
conv_dw_13(Depth Wise Conv2D)	(None,7,7,1024)	9216
conv_pw_13 (conv2D)	(None,7,7,1024)	1,048,576
global_average_pooling2d	(None,1024)	0

Table 2 Confusion matrix of SVM

86	3	0	0	0
3	124	0	0	0
0	0	53	0	0
0	0	0	140	0
0	1	0	0	139

Table 3 Classification report of SVM

Classes	Precision	Recall	F1-score	Support
Cat	0.97	0.97	0.97	89
Dog	0.97	0.98	0.97	127
Elephant	1.00	1.00	1.00	53
Monkey	1.00	1.00	1.00	140
Panda	1.00	0.99	1.00	140

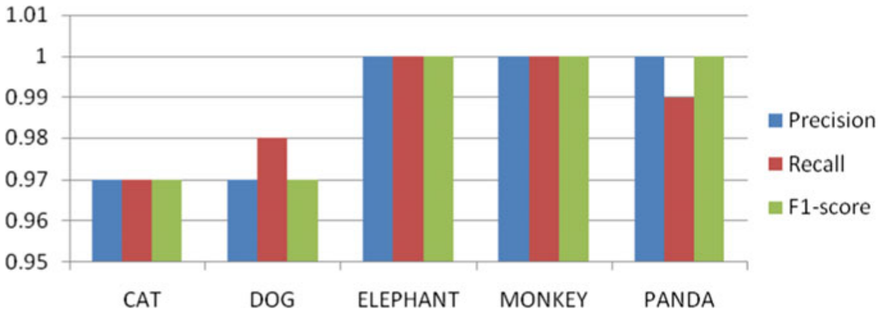


Fig. 3 Classification report of SVM

Table 4 Overall accuracy of animal classification

Accuracy			0.99	549
Macro avg	0.99	0.99	0.99	549
Weighted avg	0.99	0.99	0.99	549

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

Confusion Matrix, Classification Report, and Accuracy of SVM are shown below (Fig. 3; Table 4).

6 Conclusion

Various animal images with different backgrounds were selected for training and testing. Animal classification with MobileNet CNN architecture has achieved 99% accuracy on Animal Image Dataset. 2744 images from 5 different categories are used; out of which, 2195 images are used for training purpose. Wildlife images obtained in the field provide a difficult problem in animal classification since they can have a variety of poses, backgrounds, various lighting and climate conditions, and different viewpoints and occlusions can be done in future work.

References

1. Matuska S, et al (2014) Classification of wild animals based on SVM and local descriptors. AASRI Procedia 9:25–30
2. Manohar N, et al (2019) Convolutional neural network with SVM for classification of animal images. Emerging Research in Electronics, Computer Science and Technology. Springer, Singapore, 527–537
3. Nanditha D, Manohar N (2020) Classification of animals using toy images. In: 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE
4. Suryawanshi S, Jogdande V, Mane A, Animal classification using deep learning
5. Fujimori S, Ishikawa T, Watanabe H, Animal behavior classification using DeepLabCut. In: 2020 IEEE 9th Global Conference on Consumer Electronics (GCCE). IEEE
6. Noor A, et al (2020) Automated sheep facial expression classification using deep transfer learning. Computers Electronics Agriculture 175:105528
7. Miao Z, et al (2019) Insights and approaches using deep learning to classify wildlife. Scientific Reports 9(1):1–9
8. Zhu Y, Duan J, Tong W (2021) Animal fiber imagery classification using a combination of random forest and deep learning methods. J Eng Fibers Fabr, 16. 15589250211009333
9. .Lin L, Link Y (2020) Household animals classification using deep learning
10. Chen G, et al (2014) Deep convolutional neural network based species recognition for wild animal monitoring. In: 2014 IEEE International Conference on Image Processing (ICIP)
11. Mandal SN, et al (2020) InceptGI: a ConvNet-based classification model for identifying goat breeds in India. J the Institution of Engineers (India): Series B 101(5)573–584
12. Cao Z, et al (2018) Marine animal classification with correntropy-loss-based multiview learning. IEEE J Oceanic Engineering 44(4):1116–1129
13. Ardovini A, Cinque L, Sangineto E (2008) Identifying elephant photos by multi-curve matching. Pattern Recogn 41(6):1867–2187
14. Zeppelzauer M (2013) Automated detection of elephants in wildlife video. EURASIP J Image and Video Processing 2013(1):1–23
15. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 25:1097–1105
16. Shobha RN, et al (2018) Deformed character recognition using convolutional neural networks. Int J Engineering Tech 7(3):1599–1604
17. Adedeji O, Wang Z (2019) Intelligent waste classification system using deep learning convolutional neural network. Procedia Manufacturing 35:607–612
18. Peng Y, et al (2019) A hybrid convolutional neural network for intelligent wear particle classification. Tribology Int 138:166–173
19. Li J, et al (2018) Automatic classification of fetal heart rate based on convolutional neural network. IEEE Internet of Things J 6(2):1394–1401
20. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto Mand Adam H (2017) Mobilenets: Efficient convolutional neural networks for mobile vision applications. Retrieved from http://arxiv.org/abs/1704.04861
21. Biswas D, Su H, Wang C, Stevanovic A, Wang W (2018) An automatic traffic density estimation using Single Shot Detection (SSD) and MobileNet-SSD. Phys Chem Earth. https://doi.org/10.1016/j.pce.2018.12.001
22. Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. Retrieved from http://arxiv.org/abs/1502.03167

23. https://www.kaggle.com/piyushkumar18/animal; https://www.kaggle.com/piyushkumar18/animal-image-classification-dataset-image-classification-dataset
24. Wang W, et al (2020) A novel image classification approach via dense-MobileNet models. Mob Inf Syst
25. Loukadakis M, Cano J, O'Boyle M (2018) Accelerating deep neural networks on low power heterogeneous architectures

Performance Analysis of Different Matrices-Based SLM Techniques for Reducing PAPR in OFDM Systems

Sanjana Prasad and S. Arun

Abstract High PAPR will lead to inefficiency of the wireless communication system. It also leads to the reduction of Signal to Quantization Noise Ratio (SQNR) of A/D converters and D/A converters. This, in turn, will reduce the power amplifier efficiency used in the transmitter end. The purpose of designing phase sequences based on Hessenberg matrices for Selective Mapping technique is for minimizing the value of peak to average power ratio in wireless communication. Hessenberg matrix-based Selective Mapping technique is designed to decrease peak to average power ratio found in wireless systems. Various matrices such as Binomial, Greatest Common Divisor, Toeplitz, Riemann, Kac-Murdoc-Szego are compared with the proposed scheme. By comparing it was found that the Hessenberg matrix-based SLM technique outperformed all the other existing schemes in the literature by providing a reduction of more than 90%. The results were simulated using MATLAB 2016A Software for various modulation schemes and sub-carriers. The proposed scheme provides a significant PAPR reduction value over the other conventional techniques proposed in the literature such as Binomial, GCD, Hadamard, Hilbert, Random, Lehmer, and Riemann phase sequences as well as over Conventional OFDM and SLM techniques. The proposed Hessenberg based SLM scheme has various advantages such as faster processing of the data and lesser requirement of memory. Since they do not perform low-level operations, execution time is also less in the proposed scheme leading to higher computational efficiency. It is a significant technique which when employed to the OFDM systems will give a great reduction in PAPR value. By employing this novel method, the efficiency of OFDM systems is increased.

S. Prasad (✉)
Department of Electronics and Communication Engineering, HKBK College of Engineering, Bangalore 45, India
e-mail: sanjanaprasad@hkbk.edu.in

S. Arun
Department of Electronics and Communication Engineering, CMR Institute of Technology, Bangalore 37, India
e-mail: arun.s@cmrit.ac.in

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_26

Keywords PAPR reduction · SLM technique · Efficiency · Conventional OFDM · Hessenberg matrices

1 Introduction

Modern wireless communication systems employ OFDM systems to produce high data rate communication systems. Quadrature Amplitude Modulation (QAM) when combined with Frequency Division Multiplexing (FDM) in OFDM systems provides low latency in 5G systems. 5G systems has its own edge when it comes to high data transmission rates, latency, number of connected devices and bandwidth. 5G suffers from a serious drawback such as high PAPR. Various techniques are introduced in the literature to minimize the issue of PAPR.

Orthogonal Frequency Division Multiplexing (OFDM) is a technique used for enciphering binary data on various sub-carrier frequencies. OFDM is employed in various use-cases such as wideband digital communication, Digital Television (DTV) and Audio Broadcasting (DAB), LTE (Long term Evolution), and LTE-Advanced (LTE-A). OFDM employs orthogonal sub-carrier signals with overlapping spectra to transmit data. Good Spectral efficiency, insensitivity with narrowband co-channel interference, fading and inter-symbol interference (ISI) and poor sensitivity to time synchronization errors are some of the advantages of 4G systems.

Instantaneous power becoming larger than the average power leads to high PAPR as well as coherent summing up of modulated sub-carriers. When the signal is transmitted over a non-linear high power amplifier (HPA), it will result in degradation in performance. High Power Amplifier causes non-linearity, out-of-band radiation and distortion.

PAPR can be expressed as the ratio of the largest instantaneous peak to the average power calculated per OFDM symbol. PAPR for a discrete time signal can be given as

$$\mathrm{PAPR}(x[n]) = \max_{0 \le n \le N-1} \frac{|x(n)|^2}{E\left[|x(n)|^2\right]} \tag{1}$$

where $E[.]$ is the expectation operator.

The fluctuations in the envelope are quantified by PAPR and determination of PAPR value is by using Complementary Cumulative Distributive Function (CCDF). Within the OFDM symbol, the likelihood of the probability of the OFDM's signal envelope surpassing a particular PAPR threshold can be given by

$$\mathrm{CCDF}\left[\mathrm{PAPR}\left(x^n(t)\right)\right] = \mathrm{prob}\left[\mathrm{PAPR}\left(x^n(t) \succ \delta\right)\right] \tag{2}$$

where δ—threshold level and $\mathrm{PAPR}(x^n(t))$—PAPR of the nth OFDM symbol.

Many schemes were introduced in various existing literatures for overcoming the drawback of PAPR in wireless systems. SLM technique employs numerous combinations of the generated OFDM signal and transmits the signal having least PAPR [1].

The composition of the article is done as follows. Section 2 briefs the existing techniques using SLM technique with various phase sequences which were employed for reducing PAPR. Section 3 briefs regarding the Conventional SLM technique, role of phase sequences along with the generation of various matrices. Section 4 describes about the Hessenberg matrices as well as Proposed SLM-Hessenberg scheme. Simulation results and discussion are detailed in Sect. 5 and the concluding remarks are given in Sect. 6.

2 Literature Survey

The authors proposed SLM technique for the minimization of PAPR value in multi-carrier modulation systems [2]. Newmann, Monomial phase sequences were employed for overcoming the drawback of PAPR in SLM- OFDM systems [3, 4]. In these methods, a significant decrease in complexity was achieved at the transmitter side along with the reduction of PAPR but with slight degradation in the symbol error rate (SER) performance [5] and the outcomes were similar to that of the SLM's PAPR reduction scheme. Later, block codes were employed to decrease the PAPR in OFDM systems [6].

M-ary Chaotic sequences for minimizing the PAPR without the need for SI was introduced [7] Improved SLM using Gold/Hadamard codes had the benefits of enhanced spectrum efficiency and reduced SI transmission [8]. Riemann sequences along with the DCT transform was introduced by Pepin et al. for diminishing the PAPR [9]. Choosing proper phase sequences in a random manner was avoided, and decoding of receiver was simplified. Fibonacci-Binary (FB) sequences as an alternate type of binary sequences derived from Fibonacci series for lessening value of the PAPR [10]. This scheme worked well in terms of PAPR reduction thus consuming less computing power. Pseudo-random interferometry phase sequences for SLM technique were used to minimize PAPR value in MC-CDMA systems [11]. A Centering phase sequence matrix for improving MIMO-OFDM systems PAPR efficiency [12].

Modified Chu sequences, Hadamard matrices were also proposed for overcoming the PAPR issue [13, 14]. Lehmer sequences without the use of side information in order to lessen PAPR [15]. GCD matrices as step sequences to minimize PAPR and computational overhead in OFDM systems using the SLM technique [16]. A lightweight PAPR-GCD matrix-based scheme was employed for reducing PAPR and computational complexity while also providing higher throughput due to side information free transmission [17].

3 Selective Mapping Technique

Selective mapping employs technique where the phase rotated data of the less PAPR are transmitted. This technique involves the generation of OFDM symbols $x_{(m)}$, $0 \leq m \leq M - 1$ of length N. The generated data represents the information alike to that of the conventional symbol, later transmits signal with minimum PAPR. The number of candidate signals and the type of phase sequences play a major role in the determining PAPR value. SLM technique offers the advantages of the absence of power increase and bit error rate increase [12].

3.1 Role of Phase Sequences

Phase sequences/Phase rotations can be defined as the order in which the voltage waveforms reach their respective peaks. In SLM-OFDM systems, frequency domain OFDM signals are multiplied with different phase sequences and signal having lowest PAPR value is transmitted. Best performance is achieved when we choose the Phase sequences with minimum correlation between alternative OFDM symbols. For an SLM-OFDM system, a good phase sequence set has to be selected in order to achieve good PAPR reduction as well as to reduce the computational complexity [2].

3.2 Generation of Phase Sequences

In the Conventional SLM technique, phase sequences are generated in a random manner using phase sequences $\{\pm 1, \pm j\}$. In this work, we have used various special matrices such as Binomial, Greatest common divisor (GCD), Riemann, Hadamard, Hilbert, Random, Lehmer and Hessenberg matrices as phase sequences which will be multiplied with the data blocks. A study of PAPR reduction performances of various special matrices on SLM technique has also been carried out.

3.2.1 Generation of Phase Sequences

A binomial matrix can be defined as the multiple of an involutory matrix. An Involutory matrix can be defined as a $n \times n$ matrix that is its own inverse, which means it has to satisfy the criteria $A^2 = I$. The Binomial matrix (B) (4×4) given as

$$B = \begin{bmatrix} 1 & 3 & 3 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 3 & -1 \end{bmatrix} \tag{3}$$

3.2.2 GCD Matrix

GCD matrices are symmetric, non-singular, and non-negative definite in nature. Consider $T = \{c_1, c_2, c_3, ..., c_n\}$ to be a set of different real integers. $[T] = \left(c_{ij}\right)$ where $t_{ij} = \left(c_i, d_j\right)$, the largest common divisor of c_i and d_j.

The GCD matrix (G) (4 X 4) can be given as

$$G = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 3 & 1 \\ 1 & 2 & 1 & 4 \end{bmatrix} \tag{4}$$

3.2.3 Riemann Matrix

A $n \times n$ matrix associated with the Riemann hypothesis [18]. Riemann matrix (R) (4×4) can be given as

$$R = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & 2 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & 1 & 4 \end{bmatrix} \tag{5}$$

3.2.4 Hadamard Matrix

The values are ± 1 and the rows are mutually orthogonal in this matrix. The pair of rows represents two perpendicular vectors. The Hadamard matrix (4×4) can be given as

$$H = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \end{bmatrix} \tag{6}$$

3.2.5 Hilbert Matrix

Hilbert matrices is a square matrix with unit fractions as entries.

$$H_{ij} = \frac{1}{i + j - 1} \tag{7}$$

The Hilbert matrix (4 × 4) can be given as

$$Hb = \begin{bmatrix} 1 & ^1/_2 & ^1/_3 & ^1/_4 \\ ^1/_2 & ^1/_3 & ^1/_4 & ^1/_5 \\ ^1/_3 & ^1/_4 & ^1/_5 & ^1/_6 \\ ^1/_4 & ^1/_5 & ^1/_6 & ^1/_7 \end{bmatrix} \tag{8}$$

3.2.6 Random Matrix

A matrix in which some or all elements are random variables. The Random matrix of order 4 can be given by

$$R = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & -1 & 1 & 0 \\ -1 & -1 & -1 & 1 \\ 1 & 0 & 1 & -1 \end{bmatrix} \tag{9}$$

3.2.7 Lehmer Matrix

A Lehmer matrix is the constant symmetric matrix given by

$$L = \begin{cases} ^k/_l, \, l \geq k \\ ^l/_k, \, k < l \end{cases} \tag{10}$$

$$L_{ij} = \frac{\min(i, j)}{max(i, j)} \tag{11}$$

The Lehmer matrix (4 × 4) can be given by

$$L = \begin{bmatrix} 1.0 & 0.5 & 0.333 & 0.25 \\ 0.5 & 1.0 & 0.6667 & 0.5 \\ 0.33 & 0.6667 & 1.0 & 0.75 \\ 0.25 & 0.5 & 0.75 & 1.0 \end{bmatrix} \tag{12}$$

3.2.8 Kac-Murdock-Szego Toeplitz Matrix

The Kac-Murdock-Szego Toeplitz matrix (K) of order 4 can be given as [19]

$$K = \begin{bmatrix} 1 & 0.5 & 0.25 & 0.125 \\ 0.5 & 1.0 & 0.5 & 0.25 \\ 0.25 & 0.5 & 1.0 & 0.5 \\ 0.125 & 0.25 & 0.5 & 1.0 \end{bmatrix} \quad (13)$$

4 Hessenberg Matrices

Hessenberg reduction involves the process of transforming a full matrix to Hessenberg form by means of orthogonal similarity transformation. The matrix $H \in R^{n \times n}$ is said to be a *m-Hessenberg* matrix, $m < n$ if $H_{ij} = 0$ for all $i, j = 1, 2, \ldots n$ such that $i > j + m$.nAny matrix $A \in R^{n \times n}$ can be expressed as $A = QHQ^T$A = QHQT, where Q is orthogonal and H is *m-Hessenberg* of A. A Hessenberg structure is the resultant of the Arnoldi algorithm and is often used in areas requiring less computational control and storage.

4.1 Orthogonal Upper Hessenberg Matrices

Sparse matrices are matrices in which majority of the elements in the matrix are zero. The sparsity of the matrix is defined as the quantity of void elements to the total number of elements. Sparse matrices can be reduced in size and need reduced storage. Sparse matrices have higher computational efficiency. The sparse matrices do not require low-level arithmetic operations, such as zero-adds. This, in turn, leads to dramatic improvement in execution time. An upper Hessenberg matrix is a sparse matrix having zero entries below the first sub-diagonal. The computer storage requirement for variables of a full matrix is given by $2n^2 + 4n$ and that of a upper Hessenberg matrix is given by $n^2 + 2n$. Hence, the storage requirement is reduced by half in the upper Hessenberg matrix as the number of elements is equal to one column, sub-diagonal elements of unity and rest of the values are zero. This, in turn, leads to faster processing of the data and lesser requirement of memory.

Upper triangular matrix is expressed as a square matrix in which all the values below the diagonal main is zero. An upper Hessenberg matrix is defined as the upper triangular matrix with an extra sub-diagonal. The matrix $H \in R^{n \times n}$ is said to be upper Hessenberg if and only if $H_{ij} = 0$Hij $= 0$ for all $i - 2 \geq j$. The $n \times n$ matrix is reduced to Hessenberg form by employing orthogonal similarity transformation

$$H = Q^T A Q = Q_{n-2}^T \ldots Q_2^T Q_1^T A Q_1 Q_2 Q_3 \ldots Q_{n-2} \tag{14}$$

where A—$n \times n$ real non-symmetric matrix, Q—orthogonal matrix, H—upper Hessenberg matrix and QTQ = I. The multiplication of $(n-2)$ orthogonal matrices $Q_1 Q_2 Q_3 \ldots Q_{n-2}$ calculated along with *n-2* stages of orthogonal factorization will give the value of Q. Block version of this is reduced to block upper Hessenberg form.

The $n \times n$ matrix A is divided as N × N with a uniform b block size.

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \tag{15}$$

where $A_{11} = R^{b \times b}$, $A_{12} = R^{b \times (n-b)}$, $A_{21} = R^{(n-b) \times b}$, $A_{22} = R^{(n-b) \times (n-b)}$, and $n = bN$.

In QR factorization, $A_{21} = \tilde{Q}_1 R_1$ and that $\tilde{Q}_1 = I + W_1 Y_1^T$, an orthogonal matrix of the WY form, where W and Y are $(n-b) \times n$ matrices. The upper Hessenberg matrix after several block reduction steps can be obtained in the following form

$$H = Q_T A Q \tag{16}$$

$$H = \begin{bmatrix} H_{11} & H_{12} & \ldots & \ldots & H_{1N} \\ H_{21} & H_{22} & \ldots & \ldots & H_{2N} \\ 0 & H_{32} & \ldots & \ldots & H_{3N} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & H_{N,N-1} & H_{NN} \end{bmatrix} \tag{17}$$

where each H_{ij} is a b × b matrix and $Q = Q_1 Q_2 \ldots\ldots Q_{N-1}$ with each Q_k in WY form (Horn 2012).

Some of the properties of Hessenberg matrix are listed below.

- Hessenberg matrix multiplied with a triangular matrix will lead to Hessenberg.
- Upper Hessenberg (A) and upper triangular (T) will lead to upper Hessenberg.
- Hessenberg matrices are square matrices that have sub-diagonal rank as one.

Random, orthogonal, upper Hessenberg matrices are sparse matrices in which most of the elements are zero-valued. In the sparse matrices, the memory requirement is less since only the non-zero values are stored. Also, computational time is reduced as it eliminates operations that are performed on zero elements [20].

4.2 Proposed Hessenberg Matrix-Based SLM Technique

The proposed scheme uses various phase sequences as input to the SLM technique for minimizing value of PAPR.

Step 1: Calculating the symbols to be transmitted (X), determining the modulation order (M), sub-carriers (N), and candidate phase sequences (C). The data symbols given as input are matched with constellation points M-PSK or M-QAM to give symbol sequences. Partitioned symbols are converted into various chunks of size N.

Step 2: Various phase sequences such as Binomial, Greatest Common divisor (GCD), Toeplitz, Riemann, KMS Toeplitz, and Hessenberg matrices are generated.

Step 3: The product of Data block $X = \left[X_0, X_1, X_2, \ldots X_{N-1}\right]$ and different Hessenberg phase sequences $He^{(u)} = \left[He_0^{(u)}, He_1^{(u)}, \ldots\ldots He_{N-1}^{(u)}\right]$ is obtained.

Step 4: Distinct data blocks $X^{(u)} = \left[X_0^{(u)}, X_1^{(u)}, \ldots.X_{N-1}^{(u)}\right]^T$ are multiplied with U different phase sequences $X_n^{(u)} = X_n.He_n^{(u)}$ followed by IFFT operation to extract the time domain signals from frequency domain signals which can be shown as shown $x^{(u)} = IFFT\left\{X^{(u)}\right\}$ where $u = 1, 2, \ldots.U$. For each time domain signal, PAPR is evaluated.

Step 5: Among various time domain signals obtained, the signal having the smallest PAPR signal is transmitted.

Figure 1 depicts the SLM-Hessenberg technique.

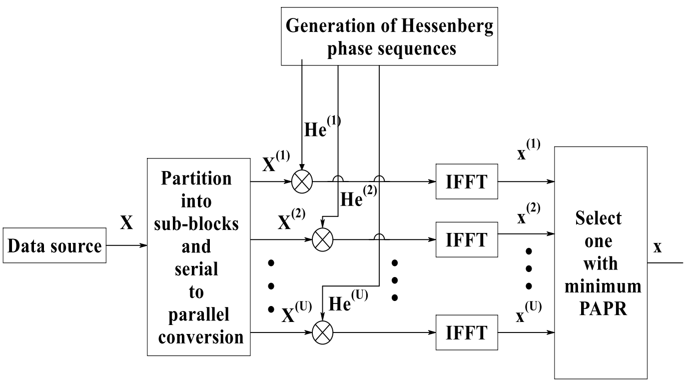


Fig. 1 Block diagram of SLM-Hessenberg scheme

5 Simulation Results and Discussion

Simulations were used to evaluate the PAPR performances of SLM when applied with various phase sequences. The simulation parameters are listed in Table 1. The phase weighting factors used are given by $\{\pm 1, \pm j\}$. Phase sequences and candidate phase sequences were considered to be 64, 128, and 256. The sub-blocks and phase weighting factors considered to be 4. The matrices used as phase sequences to the Selective mapping are Binomial, Greatest Common Divisor, Riemann, Toeplitz, Hadamard, KMS Toeplitz, and Hessenberg.

For BPSK modulation scheme, for $N = 64$, the scheme has a decrease of 10.53 dB, 5.21 dB, 3.14 dB, 2.41 dB, 2.12 dB, 1.29 dB, and 1.57 dB over C-OFDM, C-SLM, SLM-B, SLM-GCD, SLM-T, SLM-R, SLM-KMS schemes, respectively (Table 2; Fig. 2). A PAPR reduction of 11.12 dB, 5.49 dB, 4.16 dB, 2.33 dB, 2.892 dB, 1.65 dB,

Table 1 Simulation parameters

Parameter	Value
Symbols transmitted	10,000
Candidate phase sequences	64, 128, 256
Sub-carriers	64, 128, 256
CCDF	10^{-4}
Sub-blocks	4
Phase weighting factors	4
Phase sequences	Binomial, Greatest Common Divisor, Riemann, Toeplitz, KMS Toeplitz, and Hessenberg

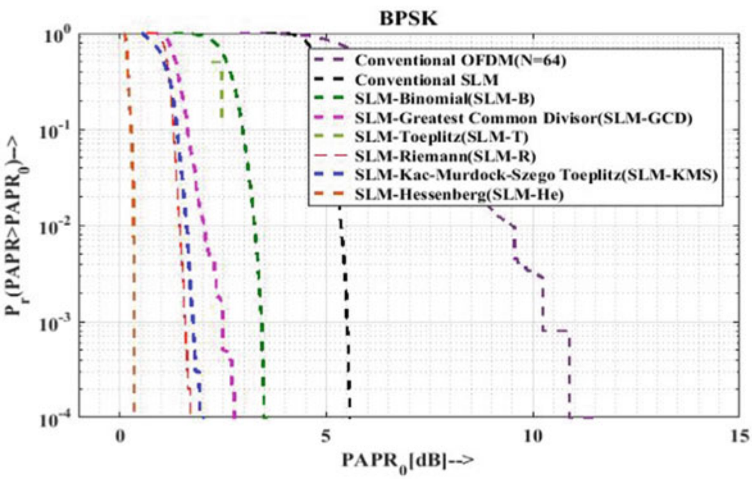


Fig. 2 CCDF comparison of various schemes for BPSK modulation and $N = 64$

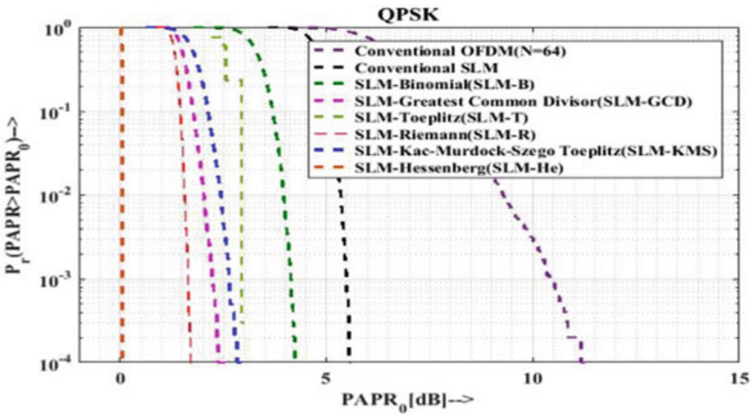


Fig. 3 CCDF comparison of various schemes for QPSK modulation and $N = 64$

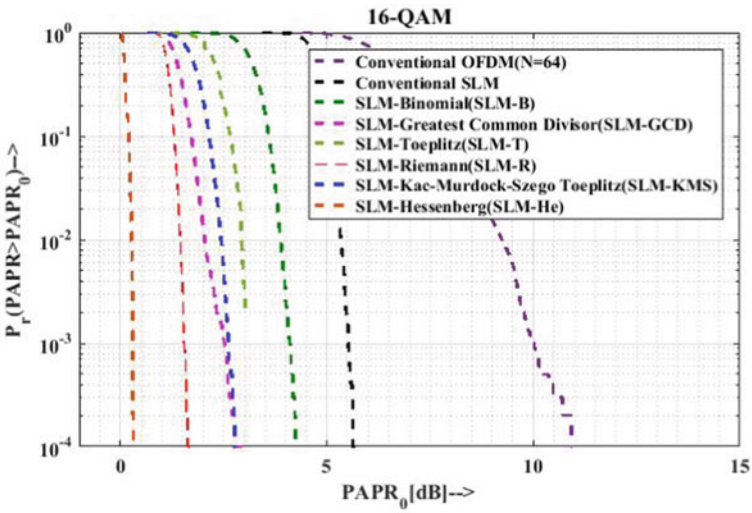


Fig. 4 CCDF comparison of various schemes for 16-QAM modulation and $N = 64$

2.78 dB is achieved over Conventional OFDM, Conventional SLM, SLM-B, SLM-GCD, SLM-T, SLM-R, SLM-KMS schemes, respectively, when QPSK modulation scheme, $N = 64$ is employed (Table 2; Fig. 3).

For $N = 64$ and 16-QAM modulation scheme, the proposed scheme gives a PAPR reduction of 3.92 dB, 2.43 dB, 2.67 dB, 1.28 dB, and 2.43 dB over SLM-B, SLM-GCD, SLM-T, SLM-R, and SLM-KMS schemes, respectively, when 16-QAM and $N = 64$ (Table 2; Fig. 4). The proposed scheme has an improvement of over 3.48 dB, 1.91 dB, 2.34 dB, 0.95 dB, and 2.19 dB when the modulation scheme when $N =$ 64 and 64-QAM modulation (Table 2; Fig. 5). In BPSK modulation scheme with

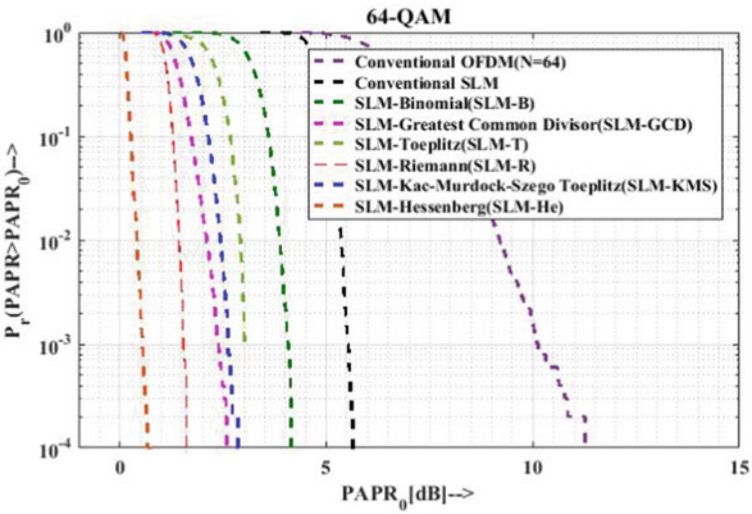


Fig. 5 CCDF comparison of various schemes for 64-QAM modulation and $N = 64$

Table 2 PAPR reduction values of different schemes when $N = 64$

Techniques $N = 64$	16-QAM	64-QAM	BPSK	QPSK
Conventional OFDM	10.93	11.26	10.88	11.17
Conventional SLM	5.626	5.636	5.567	5.546
SLM-Binomial	4.238	4.146	3.488	4.21
SLM-GCD	2.752	2.583	2.762	2.384
SLM-Toeplitz	3.006	3.009	2.475	2.942
SLM-Riemann	1.597	1.618	1.644	1.703
SLM-KMS	2.752	2.862	1.926	2.83
SLM-Hessenberg	0.3127	0.6633	0.3479	0.05

$N = 128$, the proposed scheme provides a reduction of about 5.84 dB, 11.62 dB, 3.37 dB, 1.59 dB, 2.34 dB, 1.12 dB, and 1.48 dB over Conventional SLM, Conventional OFDM, SLM-B, SLM-GCD, SLM-T, SLM-R, SLM-KMS schemes, respectively (Table 3; Fig. 6). PAPR reduction of 11.02 dB, 6.01 dB, 4.332 dB, 1.724 dB, 2.886 dB, 1.24 dB, and 2.28 dB is achieved over Conventional OFDM, Conventional SLM, SLM-B, SLM-GCD, SLM-T, SLM-R, SLM-KMS schemes, respectively, when QPSK modulation scheme is applied (Table 3; Fig. 7). For 16-QAM, the proposed scheme gives a PAPR reduction of 4.06 dB, 1.50 dB, 2.707 dB, 0.92 dB, and 2.131 dB over SLM-B, SLM-GCD, SLM-T, SLM-R, and SLM-KMS schemes, respectively (Table 3; Fig. 8). For 64-QAM modulation, the proposed scheme has an improvement of over 3.46 dB, 1.04 dB, 2.17 dB, 0.40 dB, and 1.70 dB over the SLM-B, SLM-GCD, SLM-T, SLM-R, and SLM-KMS schemes (Table 3; Fig. 9).

Table 3 PAPR reduction values of different schemes when $N = 128$

Technique ($N = 128$)	16-QAM	64-QAM	BPSK	QPSK
Conventional OFDM	10.91	11.32	11.8	11.08
Conventional SLM	6.011	6.046	6.02	6.064
SLM-Binomial	4.321	4.224	3.551	4.383
SLM-GCD	1.76	1.81	1.77	1.775
SLM-Toeplitz	2.962	2.931	2.516	2.937
SLM-Riemann	1.179	1.169	1.302	1.292
SLM-KMS	2.386	2.468	1.663	2.338
SLM-Hessenberg	0.255	0.7602	0.1752	0.05037

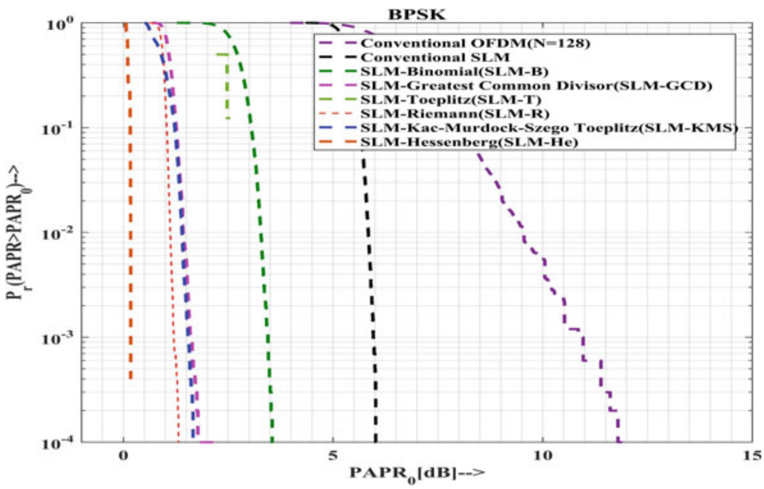


Fig. 6 CCDF comparison of various schemes for BPSK modulation and $N = 128$

In BPSK modulation scheme with $N = 256$, the proposed scheme provides a reduction of about 6.471 dB, 11.99 dB, 3.85 dB, 1.215 dB, 2.486 dB, 0.89 dB, and 1.317 dB over Conventional SLM, Conventional OFDM, SLM-B, SLM-GCD, SLM-T, SLM-R, SLM-KMS schemes, respectively (Table 4; Fig. 10) PAPR reduction of 11.55 dB, 6.44 dB, 4.58 dB, 1.205 dB, 2.49 dB, 0.95 dB, and 2.032 dB is achieved over Conventional OFDM, Conventional SLM, SLM-B, SLM-GCD, SLM-T, SLM-R, SLM-KMS schemes, respectively, when QPSK modulation scheme is applied (Table 4; Fig. 11). For 16-QAM, the proposed scheme gives a PAPR reduction of 4.32 dB, 0.91 dB, 2.42 dB, 0.58 dB, and 1.904 dB over SLM-B, SLM-GCD, SLM-T, SLM-R, and SLM-KMS schemes, respectively (Table 4; Fig. 12). For 64-QAM modulation, the proposed scheme has an improvement of over 4.51 dB, 1.12 dB, 2.62 dB, 0.784 dB, and 2.088 dB over SLM-B, SLM-GCD, SLM-T, SLM-R, and SLM-KMS schemes (Table 4; Figs. 11, 12 and 13).

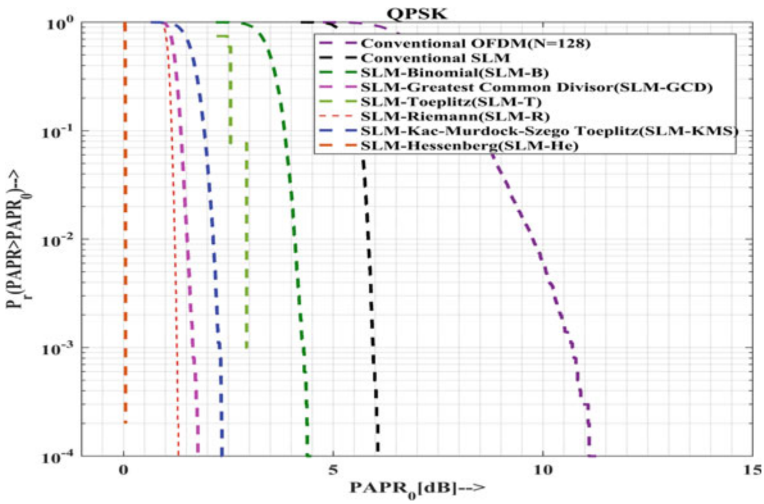


Fig. 7 CCDF comparison of various schemes for QPSK modulation and $N = 128$

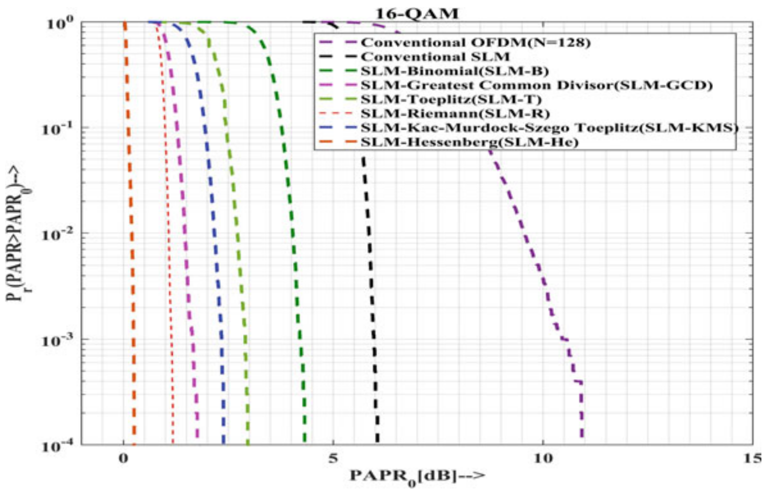


Fig. 8 CCDF comparison of various schemes for 16-QAM modulation and $N = 128$

The matrices Binomial, GCD, Riemann, Toeplitz, KMS, and Lehmer are all dense matrices consisting of mostly non-zero values. As the order of these matrices increase, the value of the elements in the matrices also increases, leading to lesser minimization of PAPR. In contrast to the above schemes, the values in the random, orthogonal, upper Hessenberg matrices do not increase with the order of the matrix and reduce PAPR to a greater extent as they are sparse in nature. In addition, it also has advantages such as lesser memory requirement as it stores only non-zero elements and higher

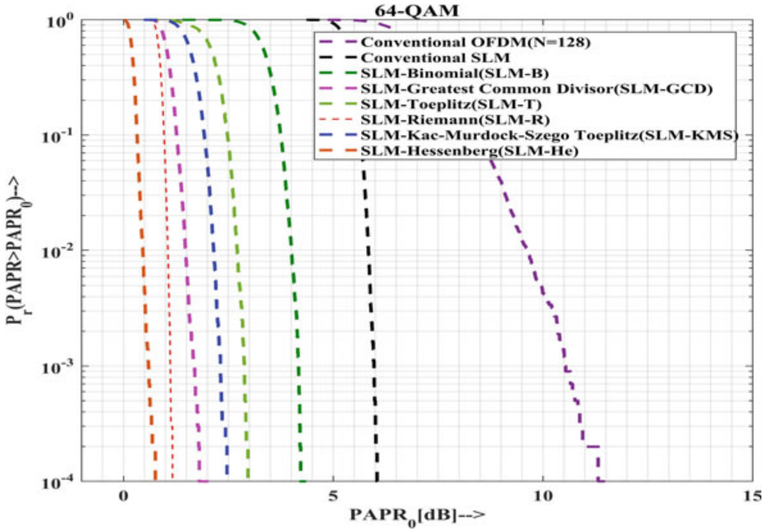


Fig. 9 CCDF comparison of various schemes for 64-QAM modulation and $N = 128$

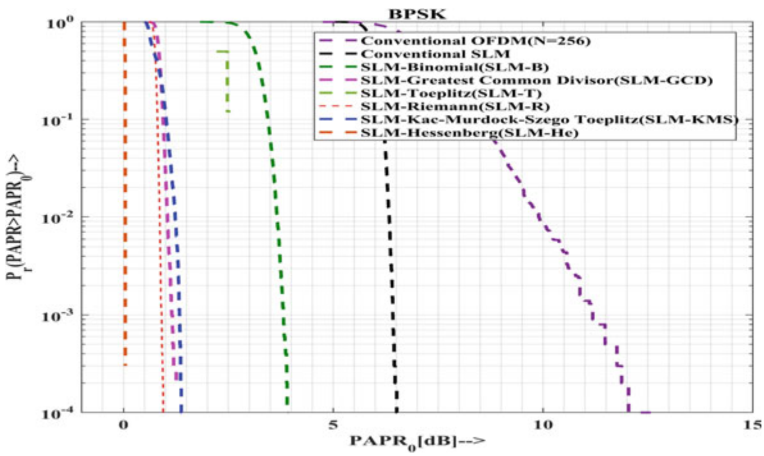


Fig. 10 CCDF comparison of various schemes for BPSK modulation and $N = 256$

computational efficiency as it does not perform low-level arithmetic operations. The proposed scheme SLM-Hessenberg is independent with respect to the modulation scheme and the number of sub-carriers (Table 5).

Table 4 PAPR reduction values of different schemes when $N = 256$

Technique ($N = 256$)	16-QAM	64-QAM	BPSK	QPSK
Conventional OFDM	11.64	11.27	12.04	11.61
Conventional SLM	6.482	6.477	6.515	6.508
SLM-Binomial	4.608	4.573	3.902	4.643
SLM-GCD	1.194	1.184	1.26	1.265
SLM-Toeplitz	2.712	2.682	2.531	2.553
SLM-Riemann	0.8681	0.8459	0.9397	1.014
SLM-KMS	2.187	2.149	1.362	2.092
SLM-Hessenberg	0.2824	0.06092	12.04	11.61

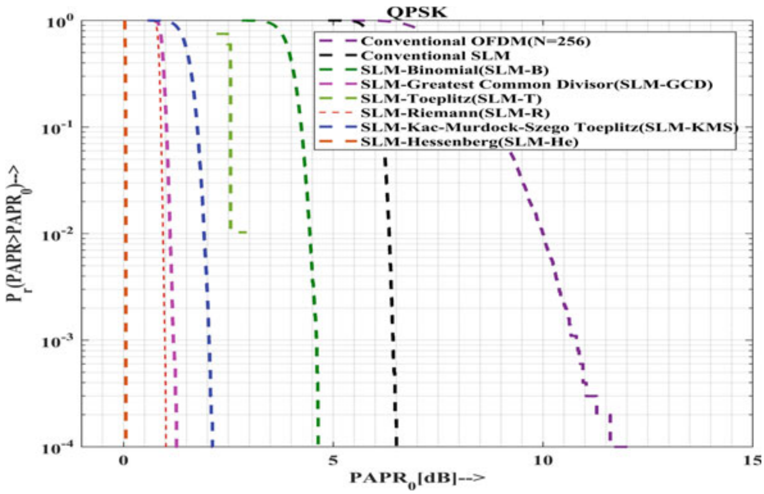


Fig. 11 CCDF comparison of various schemes for QPSK modulation and $N = 256$

6 Conclusion

In this article, the phase sequences used are Hessenberg matrices in the SLM technique for minimizing PAPR in OFDM systems. From the simulation results, the proposed SLM-Hessenberg (SLM-He) scheme is a very simple scheme to implement, as this scheme offers no difficulty in the implementation. Sparse matrices reduce storage space by assigning positions to non-zero values in the original matrix, when number of zeros is greater than non-zero ones, and this makes computations faster leading to the advantages of lower computational effort and the requirement of less storage. Analyzing the simulation results, the SLM-He scheme can be seen as one of the most effective phase sequences that can be applied to SLM technique for reducing PAPR. The proposed scheme is independent of the number of sub-carriers

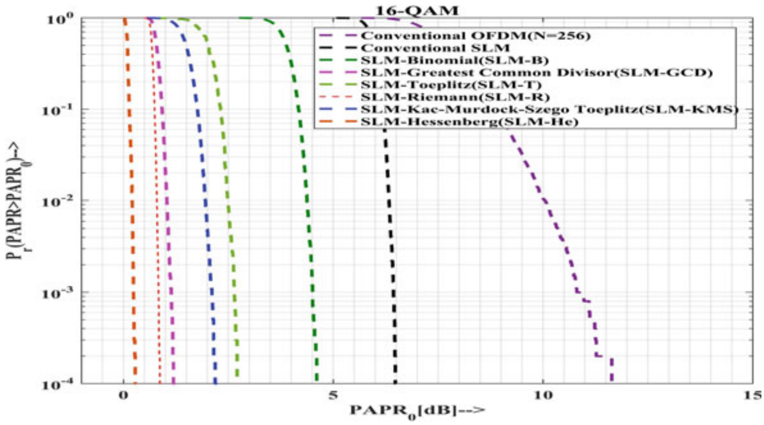


Fig. 12 CCDF comparison of various schemes for 16-QAM modulation and $N = 256$

Table 5 Comparison of the PAPR value of the proposed approach with other existing works

Technique	PAPR (dB)
Greatest common divisor	2
Riemann	1.8
Hadamard	7.8
Gold	7.5
Lehmer	6.7
Random	9
Hilbert	6.4
Circulant	8
Symmetric normalized Toeplitz + Riemann Matrices	5
Proposed SLM-Hessenberg	0.22

and the modulation type**.** The proposed scheme has various advantages such as faster processing of the data, a lesser requirement of memory and higher computational efficiency. The proposed scheme has scope in the areas of low-bandwidth applications such as Cordless phones, Bluetooth devices, Wi-Fi, near-field communication devices (NFC) and restricted frequency bands as it needs lesser power and lower processing.

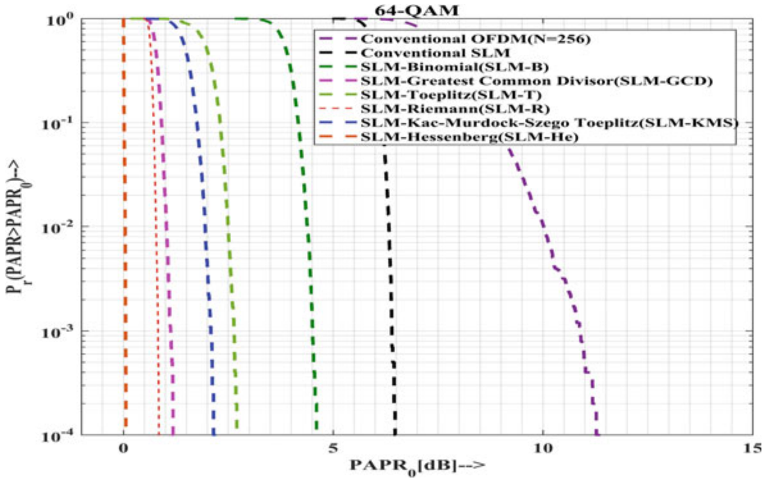


Fig. 13 CCDF comparison of various schemes for 64-QAM modulation and $N = 256$

References

1. Rahmatallah Y, Mohan S (2013) Peak-to-average power ratio reduction in OFDM systems: A survey and taxonomy. IEEE Comm Surveys Tutorials 15(4):1567–1592
2. Bauml RW, Fischer RF, Huber JB (1996) Reducing the peak-to-average power ratio of multicarrier modulation by selected mapping. Electron Lett 32(22):2056–2057
3. Zhou GT, Baxley RJ, Chen N (2004) Selected mapping with monomial phase rotations for peak-to-average power ratio reduction in OFDM. Int Conference Comm, Circuits Syst 1:66–70
4. Irukulapati NV, Chakka VK, Jain A (2009) SLM based PAPR reduction of OFDM signal using new phase sequence. Electron Lett 45(24):1231–1232
5. Heo SJ, Noh HS, No JS, Shin DJ (2007) A modified SLM scheme with low complexity for PAPR reduction of OFDM systems. IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications, 1–5
6. Kang SG (2006) The minimum PAPR code for OFDM systems. ETRI J 28(2):235–238
7. Goel A, Agrawal M, Poddar PG (2012) Mary Chaotic sequence based SLM-OFDM system for PAPR reduction without side information. Int J Electrical, Computer, Energetic, Electronic Communication Eng 6(8):783–788
8. Namitha AS, Sameer SM (2014) An improved technique to reduce peak to average power ratio in OFDM systems using Gold/Hadamard codes with selective mapping. International Conference on Signal Processing and Communications (SPCOM), Vol 22, 1–6
9. Goyoro PM, Moumouni IJ, Abouty S (2012) SLM using Riemann sequence combined with DCT transform for PAPR reduction in OFDM communication systems. World Academy of Science, Engineering Tech 6(4):313–318
10. Adegbite SA, McMeekin SG, Stewart BG (2014) Performance of Fibonacci-Binary Sequences for SLM based OFDM Systems. WSEAS Trans Comm 13:486–496
11. Wang J, Luo J, Zhang Y (2007) A new phase sequence for SLM in MC-CDMA system. In: International Conference on Wireless Communications, Networking and Mobile Computing, 938–941
12. Chanpokapaiboon T, Puttawanchai P, Suksompong P (2011) Enhancing PAPR performance of MIMO-OFDM systems using SLM technique with centering phase sequence matrix. In: Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI) Association of Thailand-Conference, 405–408

13. Sharma PK, Sharma C (2012) PAPR reduction using SLM technique with modified Chu sequences in OFDM system. MIT Int J Electron Commun Eng 2(1):23–26
14. Haque SS, Mowla MM, Hasan MM, Bain SK (2015) An algorithm for PAPR reduction by SLM technique in OFDM with hadamard matrix row factor. In: International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), 1–7
15. Vaiyamalai S, Mahesula S, Dhamodharan SK (2017) PAPR reduction in SLM–OFDM system using Lehmer sequence without explicit side information. Wireless Personal Comm 97(4):5527–5542
16. Mehata SM, Bhavani BD (2014) GCD matrix based PAPR reduction technique for OFDM system. Int J Signal Processing, Image Processing Pattern Recognition 7(5):177–186
17. Rahman MM, Rahim MS, Bhuiyan MN, Ahmed S (2015) Greatest common divisor matrix based phase sequence for PAPR reduction in OFDM system with low computational overhead. In: International Conference on Electrical & Electronic Engineering (ICEEE), 97–100
18. Jeon HB, No JS, Shin DJ (2011) A low-complexity SLM scheme using additive mapping sequences for PAPR reduction of OFDM signals. IEEE Trans Broadcast 57(4):866–875
19. Trench WF (2001) Properties of some generalizations of Kac-Murdock-Szego matrices. Contemp Math 281:233–246
20. Berry MW, Dongarra JJ, Kim Y (1995) A parallel algorithm for the reduction of a nonsymmetric matrix to block upper-Hessenberg form. Parallel Comput 21(8):1189–1211

Fake News Detection in Mainstream Media Using BERT

Ashwin Sudha Arun, V. P. Gautum Subhash, and S. Shridevi

Abstract Fake news classification is undoubtedly a challenging task that demands to be resolved in this world of media and mass communications. Despite the majority of the falsified information being strictly filtered, some of the fake news makes it into platforms of large audiences such as news and radio channels. This paper proposes a relatively simple yet reliable solution for real-time fake news classification focused on mass media sources such as news and radio channels. The proposed model focuses on the user's multiple ways to input the news from live media from which relevant data can be extracted. Then, it utilizes a natural language processing (NLP) framework called BERT to extract contextual features from the raw text data and classify the news as fake or real. Experimental results obtained reflect the effectiveness of the model in detection of fake news when the headline and supporting text are provided as an input. The proposed model achieves an accuracy of 96.5% on the test set classification into two classes—real and fake.

Keywords Fake news · Machine learning · Natural language processing · BERT

1 Introduction

The constant evolution of technology has led to the immense growth of media. Today, communications and media play a vital role in forming public opinion and bringing people's attention to various events across the globe. Platforms for sharing news and factual information are ever increasing in popularity. Along with this trend,

A. S. Arun
School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India
e-mail: ashwinsudha.arun2018@vitstudent.ac.in

V. P. G. Subhash
School of Electrical Engineering, Vellore Institute of Technology, Chennai, India
e-mail: gautumsubhash.vp2018@vitstudent.ac.in

S. Shridevi (✉)
Center for Advanced Data Science, Vellore Institute of Technology, Chennai, India
e-mail: Shridevi.S@vit.ac.in

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_27

comes the risk of spreading misinformation among its audience. Misinformation or fake news is often created by an individual or an organization with the intention of harming another person, group or community [1]. Fake news is a major concern in our society as they possess the power to sway the public opinion in their favour. Majority of this false information is aimed at affecting financial or political affairs and is targeted towards the general public. These tend to create a perception of the situation in the public minds rather than the whole truth. The media witnessed a huge rise in sharing of fake news or information during the last Presidential Elections in the US. Similar rises in fake news spread were observed during major events across various countries. Verification of news became more of a legal responsibility rather than a moral conduct to be followed by journalists and publishers. Various forms of traditional checking and verification practices are being followed by the media to restrict the leak of any false news.

Accounting of evidence and backtracking are other alternative methods. Despite such measures, false news is being introduced not just on social media platforms but also on mainstream media such as news and radio channels. This has drawn the attention of researchers as they attempted to analyse patterns using lexical and numeric methods in such data in an attempt to classify the news as genuine or falsified [2]. Today, automated systems or models are being developed to keep up with the copious amounts of information flowing in through media and other sources. These models utilize existing data to draw patterns that account for similarities and differences which can possibly determine the status of a piece of information. Various text processing algorithms are performed to extract lexical features from the raw data. Other possibilities for fake news classification include sentiment analysis, artificial intelligence and so on. Different model approaches account for content-based and context-based information [3].

The main objective of this research paper is to conduct a survey of the existing models and propose a simple and effective real-time model that is capable of detecting fake news when suitable input is provided. The novelty of this model lies in the extraction of data from its environment. It allows the user to input news which is being broadcasted live on news media sources such as live TV, radio, etc. from which necessary features are extracted. The proposed model uses deep learning to classify the news as real or fake. Since this problem involves heavy processing of text, NLP plays a key role in this model. With the proposed system, we expand the fake news detection problem to mainstream media. Along with the real-time data extraction, this model acts as a base for further optimization algorithms. The model was trained on a native dataset which found similarities in real-time data and also provided us an opportunity to evaluate its performance on actual data.

2 Related Works

X. Zhou et al. in their research paper describe the challenges and strategies occurred while classifying the news into fake or real. It describes the distinction between style-based and propagation-based fake news and the suggestive methods to detect them. Shu et al. [4] give an in-depth review of the models and datasets that can be used to tackle the problem of detecting fake news. The paper describes several approaches that can be used to identify fake news. Systems can also be implemented in a stance-based approach on social media, where a semi-supervised probabilistic model can be used to predict if a Facebook post is a hoax, based on the reaction of users to it. A credibility algorithm was also described which uses data on the author's credibility and tweet implications on Twitter. Rashkin [5] has further analysed the nature of fake news based on the content properties especially when the subject of news is politics. The objective of their research paper was to highlight the significance of NLP in the detection process. Text processing and analysis plays a vital role in this scenario as the primary data we are dealing with is largely composed of the same. A similar approach is observed in the model proposed by Karimi et al. [6] which aims to perform a multi-class classification of news comments. Oshikawa et al. [3] too explore the critical role of NLP in detection or classification of fake news. They utilized the linguistic characteristics and features such as word count, term frequency (TF-IDF) of relevant words and subjective words to sentiments that contribute.

Jain et al. [2] use a combination of SVM and Naive Bayes models resulting in higher accuracy than those which use a single model. Their data of training was obtained from the RSS feeds of various websites. Their model utilized SVM and Naive Bayes to classify the news into two labels based upon the extracted features. This machine learning approach achieved an accuracy of 93.5% on test data. Pérez-Rosas et al. [7] use a linear SVM classifier on a set of custom-made datasets. The dataset was made from crowdsourcing methods and from websites. Karimi, H. et al. use a system that has three steps to identify the degree of the fake content within the news from different sources. A CNN and LSTM network identifies and extracts the patterns in the data. The attention mechanism is used to extract features from multiple sources and classification is done to separate and group similar samples. This earns its accuracy when compared to other regular models for the same dataset. Long, Y. et al. proposed a model that takes into account the features that characterize the speaker such as profile, affiliation, etc. Even after using LSTM, it is underachieved in terms of model accuracy. Bhatt et al. [8] accounted for word embedded features such as weighted bag of word, term frequency (TF-IDF) to extract features from headline and text in order to classify them. They utilized deep neural networks as a backbone to achieve a classification accuracy of 43.82%. Han et al. [9] also suggests a rather complex solution to a problem of similar domain. It elaborates the use of graph neural networks it utilizing the various parameters of social media to detect fake information.

On the other hand, Reis et al. [10] explore the extraction of several features including textual (language, lexical, etc.), news source features (credibility of publisher, etc.) and environmental features accounting for social media statistics to build the classification model. Rashkin et al. also followed a similar approach by analysing the linguistic characteristics of the headline or body in the detection process. This resulted in evidence that the random forest and XGB model outperformed all the other classification models. Classification accuracy for both was found to be around 81%. Ensemble learning is also gaining popularity in the field of data prediction. Kaliyar et al. [11] proposed a similar model to handle the overfitting of the dataset while reducing bias and variance. This model explores a multiclass dataset along with various feature extraction—term frequency, cosine similarity, word overlap, etc., to identify certain word similarities. The gradient boosting allows the training of a few models in a gradual and sequential manner. It utilizes the gradients in the loss function to minimize bias. Thus, the proposed system achieves an accuracy of 86% in multiclass textual classification.

The papers reviewed provide us with a clear understanding of the working of the models used to identify fake news from a given dataset. Several models such as knowledge graphs, random forest, CNN with bidirectional LSTMs and neural networks have been implemented providing varying accuracy according to the dataset used. Ahmed et al. [12] make it clear that social media data and mainstream news data are very different datasets as mainstream news datasets have lesser features when compared to social media datasets. However, models using NLP can be effectively used in such cases. Most of the other models rely on repeated words or patterns and even on whether similar articles have been published before. Thus, there is potential in creating a model that can obtain real-time data to determine the veracity of the information.

3 System Model

Our model input is assumed to be raw text which conveys news information of a certain subject. This text may be a combination of the title or headline followed by the associated details. These data are obtained in real time from mainstream live media sources such as TV or radio channels. Our proposed model provides the user two options through the user can capture data as in Fig. 1.

3.1 Using Speech To Text Recognizer

For this method, we assume a scenario where the user has the means to access a TV or radio broadcast where specific news of the user's interest is being read out by a newsreader or a host. The user can utilize their recording device to record the relevant part of the news. This part of speech can be transcribed into raw text which can be pre-processed for the model. Various packages have been used and compared for the same purpose. We have opted to use the PyAudio package as it provides us the Python

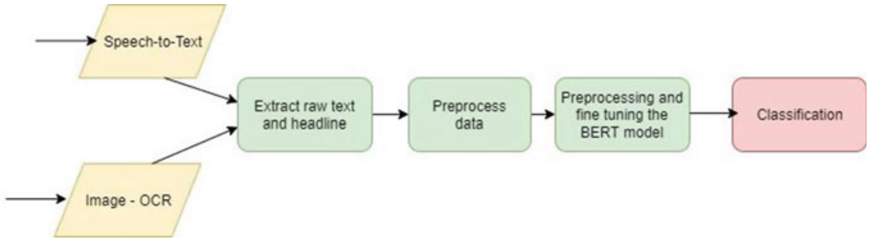


Fig. 1 Block diagram of the system model. Two forms of input—speech and image—are obtained and undergo the respective processes

bindings for PortAudio which is required for audio recording, saving and reviewing. Other features include noise suppression along with modification of stream, number of channels and so on. An associated open-source Python library PyAudio analysis also allows other procedures such feature extraction, segmentation and visualization. But since our model involves in-depth feature extraction of the test rather than the audio, we focus on recording of the speech alone. We utilize the services offered by various speech APIs to transcribe the speech that was previously recorded. Various large corporations and firms provide such interfaces for its users to perform such machine learning tasks. Using an HTTP request, these services can be accessed. However efficient they may be, they differ in terms of certain characteristics and extensibility.

3.2 Using Image OCR

Our second method of obtaining data for classification is via images captured by the user. The user is required to use their device to capture the screen of a TV news broadcast which contains the headline followed by any additional text or scroll text. This can be in the form of a well-adjusted image focusing on the TV screen or a screenshot of the broadcast on social applications or video sharing platforms such as YouTube. In order to prepare the image for text retrieval, a series of pre-processing techniques are utilized. Some of them include:

- Rotation and adjusting skewness
- Grayscale conversion
- Thresholding
- Dilation and erosion
- Noise removal

Bieniecki et al. [13] and Gupta et al. [14] proposed these techniques as a means of overcoming the limitations faced by digital cameras. This is followed by line and word detection and feature extraction to recognize the characters within the image. Character recognition works by either evaluating the lines and strokes or later

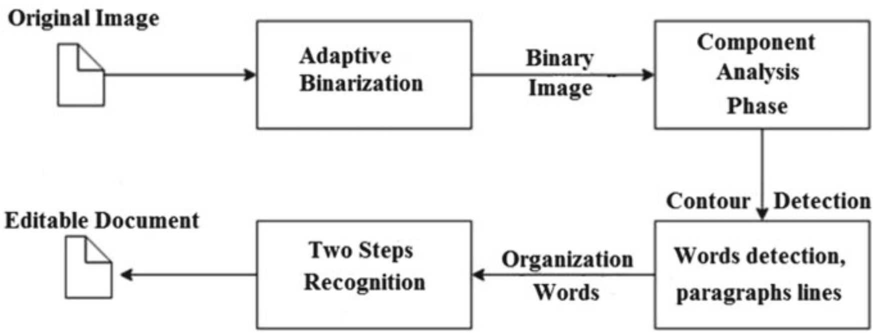


Fig. 2 Working of Tesseract [15]. This diagram shows the various steps

conjoining them to make the entire character or by identifying the character as a whole. The accuracy of recognition is improved by invoking the use of lexicons, i.e. permitting certain words in the document. It could be from a particular field or language (Fig. 2).

Tesseract uses a LSTM implementation re-designed in C++ instead of Python. LSTMs are popular when it comes to sequence learning. It detects the words by organizing text lines. Then it later processes them for recognizable words after identifying them in terms of spacing. The input image is scanned using rectangular boxes line by line and then sent to the LSTM/RNN model for classification. Ribeiro et al. [15] used PyTesseract for the same purpose. It is an implementation of Tesseract-OCR in Python that is capable of reading all popular image formats.

3.3 BERT

The pre-processing of the training data is succeeded by the phase of classification using NLP frameworks. Our objective is to establish a contextual relationship between a key or token word and its adjoining words. This helps the model gain an understanding of when the news provided is likely to be fake depending on the context rather than just content. This is achieved by BERT which stands for bidirectional encoder representations from transformers. Transformers are a deep learning framework designed for NLP problems involving sequential learning. It focuses on providing context to any part of the text and is majorly used for processes such as summarization. BERT, which is a sub-product of this framework, represents input as sub-words and learns embeddings for sub-words. BERT unlike the sequential models does not depend on previous input but based on an input attention mechanism. One of the features that make BERT ideal for our problem is the bidirectional feature. Devlin et al. [16] took into account the context learning from both sides of the targeted word, i.e. left and right. Hence, the classification process is enhanced since two-way context of the word is considered. The working of BERT is depicted in Fig. 3.

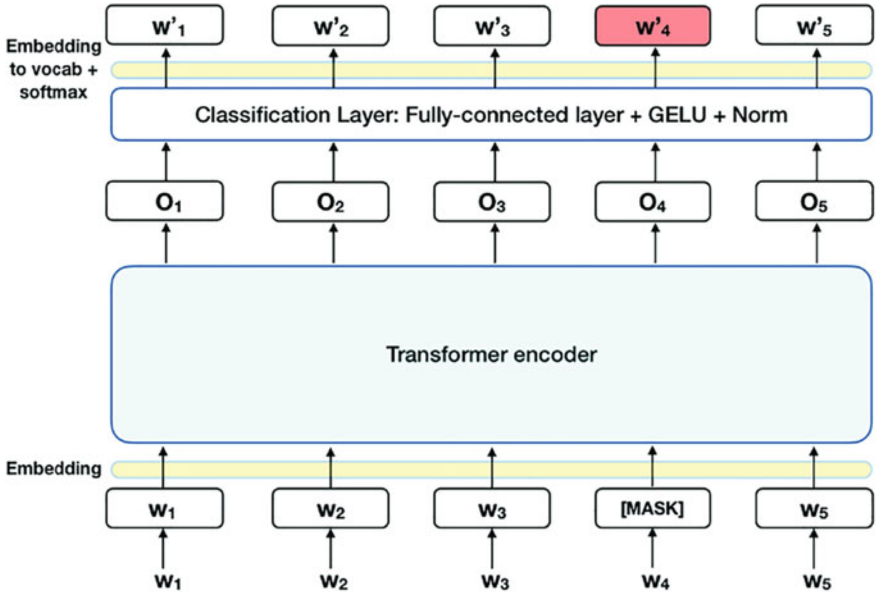


Fig. 3 Classification and working of BERT. Transformer performs the encoding process by converting individual words into word embedding vectors which are passed into attention-based encoder to obtain its context

The probability of a word '*i*' being start of the related words to a given context is given as follows

$$P_i = \frac{e^{S \cdot T_i}}{\Sigma_J e^{S \cdot T_j}}$$

S is the start vector and *T* is the output of the position in the given context. Dot product of *S* and T_i gives the probability of the target words starting at position *i* in the context, while that of E and T_i give the score of how likely the target ends at position *i*. The output is then passed through a softmax function to calculate probability.

When compared with the previous word embedded model words, a two-way sequential relationship is explored by BERT. This prevents any loss of information. BERT is primarily pre-trained on two tasks—masked language prediction and next sentence prediction. BERT is first trained on a large unlabelled text dataset and then fine-tune the model according to our specific problem. This eliminates biasing the word towards a specific meaning and instead broadens the utility or the context of its use such as Word2Vec and GloVe, BERT provides an extended knowledge of words and their respective context, rather than depending on a limited correlation between two keywords.

4 Implementation and Result

Dataset used is a collection of news classified as 'fake' or 'true'. There are about 25,000 different news in this dataset. The real news data in this dataset is obtained from websites like Reuters and other news websites. The fake news data was collected from certain websites that exist to flag the level of veracity of the news. Examples of such websites include Politifact. Most of the data is concentrated on the US elections and related politics. To train the model, a concatenation of the title and text of the news was used. Unnecessary data like the date of publication was removed. The data was collected from the US elections due. The different subjects mentioned in the dataset are shown below in Fig. 4:

The distribution of fake and real news in the dataset is shown below in Table 1.

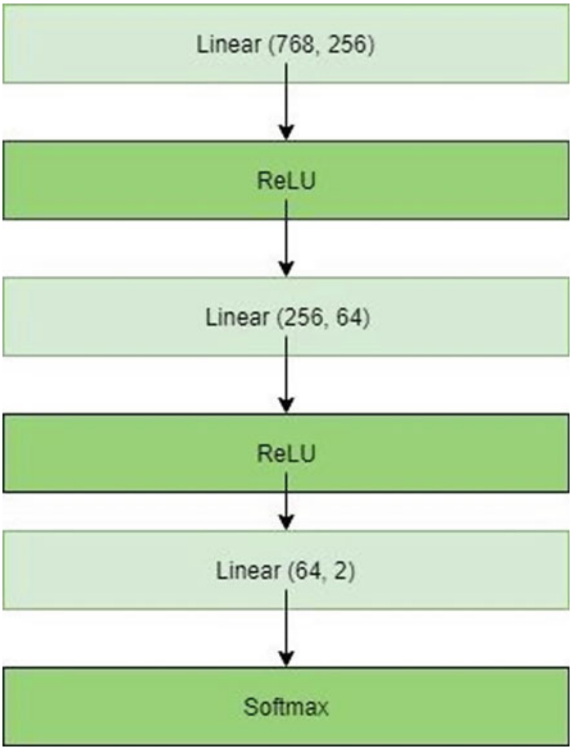


Fig. 4 CNN architecture of proposed model

Table 1 Distribution of fake and real data

Dataset	Number of instances	
	Title	Text
Real	20,826	20,826
Fake	17,903	17,903

The dataset contains 20,826 numbers of verified real news and 17,903 numbers of fake news. As mentioned before, since much of the news is based on the US politics, the popular words in it reflect the same. The dataset provides not only the text but also a detailed text of the report which can be used to train the model. Natural language toolkit (NLTK) is a collection of libraries and programmes in Python used for natural language processing in English. The use of this suite allows the user to refine the text input, allowing the creation of a model with higher accuracy. Kannan et al. [17] and Vijayarani et al. [18] used the appropriate pre-processing techniques in an attempt to save space and time complexity involved in the process.

(a) **Stop words**: NLTK is used to remove stop words. Stop words are commonly used words in a language. They carry little information. Examples include words like ‘a’, ‘the’, etc. Since they are common in all the entries of the dataset, they were removed. Other common words spoken on live media like ‘Breaking News’ was also removed.
(b) **Punctuation**: Removal of punctuation increases performance as unwanted data is removed. Punctuation holds no useful information and removing it will reduce the noise present in the system. When tokenization is done using NTLK, the sentence is split into words. The special characters will be not part of the split words. The words are then concatenated into a string.
(c) **Arranging the data**: The data is labelled and then shuffled before training. The title and text of the data are merged because, in real-life cases, the title of the story is read out and then the rest of the story. The words are also converted to lower case for higher accuracy while training the model.

5 Results

The user takes input from their device’s microphone. Google speech recognition API is used to identify the words spoken and is given as input to the classifier. Image recognition can also be used to recognize text in a given image which can be a screencap of the news headline. However, this method is less effective than voice recognition as the input data has a lower number of words. Dataset consisting of real news from reputed sources and fake news flagged by professional fact-checkers was used to train the model. The data revolved mostly around the US elections and the buzz around it due to which the test input also had to be news of the same period. The CNN model used in this project uses a combination of linear layer, ReLU and softmax layer for the final layer. Figure 5 depicts the working of CNN.

The softmax layer is used to convert a given input vector to a similar-sized vector with elements in the range of 0 and 1. The use of a softmax layer allows us to convert scores in such a way that a normalized probability distribution is obtained. Rectified linear unit (ReLU) will convert negative input values to 0. Positive input values are given out as it is. Models using ReLU are easier to train and have higher performance due to which they are a popular activation function for neural networks. Figure 6 depicts the training loss over the epochs.

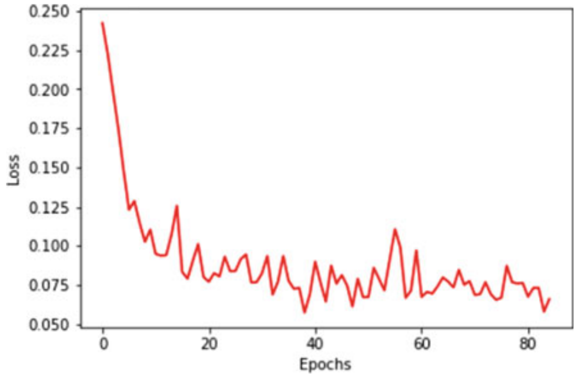


Fig. 5 Training loss plotted over the epochs

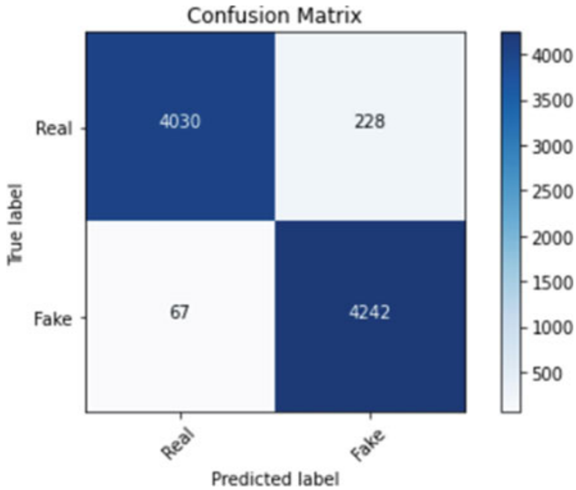


Fig. 6 Confusion matrix represents how well the model did in classifying the input as fake or real

25,700 epochs were done to train the BERT model with the training loss falling from 0.247 to 0.066 during the process. Figure 7 visualizes the loss over the epochs. On the test set whose training had 8567 epochs, the model gave an accuracy of 96.55%.

ROC AUC is an important tool for measuring performance. It shows the capability of a model to differentiate between the two classes. The closer the AUC value is to 1, the better the model is at classifying. The model used in this paper has a ROC AUC of 0.965. F1 score is the weighted average of precision and recall. Similar to AUC ROC, the closer the value is to 1, the better the model. The model used had an F1 score of 0.9655. Table 2 gives the details of the performance comparison with existing algorithms.

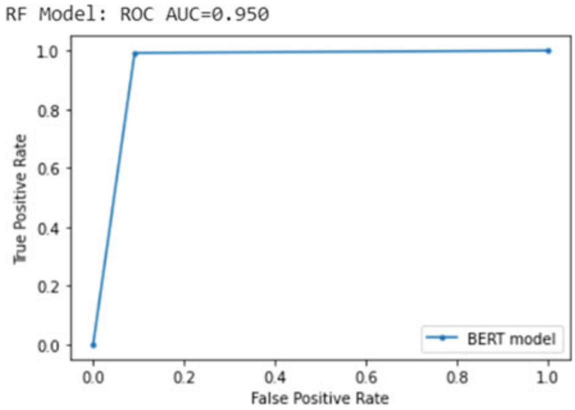


Fig. 7 ROC AUC graph shows the receiver operating characteristics curve and the area under the curve (AUC)

Table 2 Comparison analysis with existing algorithms

Implementation method	Accuracy
Logistic regression	0.63
Naïve Bayes [19]	0.754
Random forest classification	0.85
XGB	0.86
Proposed method CNN with BERT	0.965

6 Conclusion

In this paper, raw text containing news information is extracted from the environment. A CNN with BERT is trained using a dataset that has both fake and real news. Speech converted to text is saved as a file that is accessed by the model as input to the classifier. Google's speech recognition engine is used to convert speech to text. There are several other options available for speech to text with higher accuracy. The BERT classifier is providing good results when classifying, especially in cases where a large number of words are given as input. Custom inputs were given from various clips available on the internet. The model correctly identified the label but with varying degrees of certainty. Future scope in further research includes improving the accuracy of the model in more topics by including a larger and diverse dataset. The use of sophisticated speech recognition APIs will greatly improve the accuracy of speech to text conversion.

References

1. Zhou X, Zafarani R (2018) Fake news: A survey of research, detection methods, and opportunities. arXiv preprint. arXiv:1812.00315, 2
2. Jain A, Shakya A, Khatter H, Gupta AK (2019) A smart system for fake news detection using machine learning. In: 2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), vol 1, pp 1–4. IEEE
3. Oshikawa R, Qian J, Wang WY (2018) A survey on natural language processing for fake news detection. arXiv preprint arXiv:1811.00770
4. Shu K, Sliva A, Wang S, Tang J, Liu H (2017) Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter 19(1):22–36
5. Rashkin H, Choi E, Jang JY, Volkova S, Choi Y (2017) Truth of varying shades: Analyzing language in fake news and political fact-checking. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp 2931–2937
6. Karimi H, Roy P, Saba-Sadiya S, Tang J (2018) Multi-source multi-class fake news detection. In: Proceedings of the 27th international conference on computational linguistics, pp 1546–1557
7. Pérez-Rosas V, Kleinberg B, Lefevre A, Mihalcea R (2017) Automatic detection of fake news. arXiv preprint arXiv:1708.07104
8. Bhatt G, Sharma A, Sharma S, Nagpal A, Raman B, Mittal A (2018) Combining neural, statistical and external features for fake news stance identification. In: Companion Proceedings of the the Web Conference 2018, pp 1353–1357
9. Han Y, Karunasekera S, Leckie C (2020) Graph neural networks with continual learning for fake news detection from social media. arXiv preprint. arXiv:2007.03316
10. Reis JC, Correia A, Murai F, Veloso A, Benevenuto F (2019) Supervised learning for fake news detection. IEEE Intell Syst 34(2):76–81
11. Kaliyar RK, Goswami A, Narang P (2019) Multiclass fake news detection using ensemble machine learning. In: 2019 IEEE 9th International Conference on Advanced Computing (IACC), pp 103–107. IEEE
12. Ahmed H, Traore I, Saad S (2018) Detecting opinion spams and fake news using text classification. Security Privacy 1(1):e9
13. Bieniecki W, Grabowski S, Rozenberg W (2007) Image preprocessing for improving ocr accuracy. In: international conference on perspective technologies and methods in MEMS design, pp 75–80. IEEE, May
14. Gupta MR, Jacobson NP, Garcia EK (2007) OCR binarization and image pre-processing for searching historical documents. Pattern Recogn 40(2):389–397
15. Ribeiro MR, Jùlio D, Abelha V, Abelha A, Machado J (2019) A comparative study of optical character recognition in health information system. In: 2019 International Conference in Engineering Applications (ICEA), pp 1–5. IEEE
16. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
17. Kannan S, Gurusamy V, Vijayarani S, Ilamathi J, Nithya M, Kannan S, Gurusamy V (2014) Preprocessing techniques for text mining. Int J Computer Science Communication Net 5(1):7–16
18. Vijayarani S, Ilamathi MJ, Nithya M (2015) Preprocessing techniques for text mining-an overview. Int J Comp Sci Comm Net 5(1):7–16
19. Granik M, Mesyura V (2017) Fake news detection using naive Bayes classifier. In: 2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON), pp 900–903. IEEE
20. Long Y (2017) Fake news detection through multi-perspective speaker profiles. Association for Computational Linguistics

A General Approach of Plant Disease Identification and Detection from the Images Using Convolutional Neural Networks

A. Dhanalakshmi and K. Ponmozhi

Abstract Plants are really important for the planet and all living things. They provide us the oxygen which humans and other animals need to breathe. At any instances, diseases in plant can impair the plant leaves, concluding in significant cultivate losses and a significant drop in market value. In recent research it was found that disease is the most important cause for 50% of yield loses. Thus a correct methodology is needed to identify the infection earlier so that it minimizes the loss, increases the crop yield and improves the productivity. Image-based automated identification has gained much importance due to their simplicity, time consumable and provides early detection of disease. Various machine learning and deep learning technologies are being investigated for the detection of plant disease. In this work, a deep survey is made on the identification of disease in plant. This paper provides the metrics used till date for prior evaluation based on strategies for detecting and classifying plant disease symptoms. Also, a comprehensive deep learning method used for identifications of plant disease like convolutional neural network are also discussed.

Keywords Deep learning · Machine learning · CNN · Rice plant disease

1 Introduction

Plant diseases are one of several factors that cause a decline in the amount and quality of agriculture crops. In most developing countries, agriculture is the primary source of revenue. Due to political uncertainties, the agriculture sectors initiates to explore innovated ways to enhance the quantity of food. The agriculturists determines the required species of crop reliant to the farmland, the climatic conditions of the region, and other factors. Pesticides are used by farmers to govern pests, prevent from infection, and boost crop output. Diseases in plants are producing challenges for agriculturist due to minimum yield, economic losses, and industrial farming,

A. Dhanalakshmi (✉) · K. Ponmozhi
Department of Computer Applications, Kalasalingam Academy of Research and Education, Krishnankovil, Tamilnadu, India
e-mail: dhanuswathi020@gmail.com

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_28

according to a survey [1]. As a result, disease infection and rigorness are centered on identifying and detecting the correct disease.

2 Rice Plant Diseases

First, listed below are the various forms of rice plant diseases considered:

Bacterial Leaf Blight: The sample image for this type of disease is given in Fig. 1.

- *Components* of a plant where it affects: commonly affects leaves regarding the plant.
- *Symptoms Shape*: Symptoms consists of elongated lesions from the leaf tip, lesions are many inches long.
- *Lesion color*: Yellow to White as a result of the effect of bacteria (Fig. 2).

Leaf Smut: The trial image for this type of disease is presented in Fig. 3.

- *Components* of a plant that it affects: typically, affects leaves regarding the plant.

Fig. 1 Image of bacterial leaf blight disease

Fig. 2 Image of brown sport disease

Fig. 3 Sample image of leaf smut disease

- *Symptoms Models*: signs and symptoms regarding the disease are small spots scattered for the leaf in non-uniform shape.
- *Lesion color*: Dark Brown.

3 A Literature Review on Plant Disease Detection

At this time, various researchers work complete detection of plant disease Image that is using processing deep learning and machine learning models.

The authors Ferentinos [2] in 2018 various convolutional neural network models have been constructed. Overfeat, AlexNet, GoogLeNet, VGG, and AlexNetOWTBn are programs that use simple leaf photos of healthy and diseased plants to identify plant disease combinations. The image collection consists of 87,848 distinct leaf images from 58 various classes and 25 various species of plant. 99.53% of accuracy is the highest classification accuracy attained using the model VGG in convolutional neural networks.

Lu et al. [3] in 2017, using the supervised deep learning architecture, they built a model to diagnosis the wheat disease automatically. 50,000 annotated photos of healthy and sick wheat crop leaves make up the image dataset. Four distinct CNN models were built in this study. VGG-16 model recognized seven various classes of disease in wheat and attained the 97.95 percent as the highest average identification.

Lu et al. [4] in 2017 have described a rice crop disease detection approach based on deep learning. The image dataset contains 500 photographs of healthy and damaged rice leaves and stems. The CNN model developed is based on the CNN architectures AlexNet and LeNet-5 have described a deep learning-based strategy for detecting rice crop illness. There are 500 photos of healthy and unhealthy rice leaves in the image dataset. The CNN model built is based on the LeNet-5 and AlexNet CNN architectures.

Mohammed Brahimi et al. [5] in 2014, has examined a large dataset of about 14,828 images of tomato leaves infected with nine diseases. They have implemented the CNN as a learning algorithm. This is because, the raw images are directly processed by the automatic extraction of features present in CNN. This characteristic is found to be the greatest advantage of CNN. In addition to it, visualization methods are used to recognize the limit the diseased regions in leaf. The attained outcomes are inspiring with 99.18% of accuracy. These results show that this model could be practically implemented. This will be much helpful for farmers to protect tomato against disease.

Ramesh et al. [6] in 2019, the paddy leaf disease is classified by means of optimized deep neural network with java, Optimization Algorithm (DNN_JOA). By exploring this method in the post processing step, a feedback loop is created. Based on the process various highest accuracy were attained. 98.9% was achieved for blast, 95.78% accuracy for bacterial blight. Similarly, 92% precision is got for the sheath rot, 94% for the brown spot and 90.57% for the normal leaf image.

Hence, the deep learning algorithm discussed in the literature found to be significant in agricultural production, intelligent agriculture, and ecological protection.

4 Deep Learning Methods Used in Identifying Disease

Deep learning is a subset of machine learning. This utilizes a model of computing which is much stimulated by the structure and functioning of the brain. Among the different methods of recognizing diseased leaf in a plant, few are discussed below:

I. **Computer Vision and Pattern Recognition**

Artificial intelligence that instructs computers to interpret and analyze the visual world is called as Computer Vision [7]. Machines can properly detect and categorize items using movies, digital photos from digital camera or mobile cameras, and by various models. Various types of computer vision methods are available and they are employed in several ways namely:

Image Segmentation is popularly known for image partitioning into multiple regions which could be inspected separately [8].

Object Detection, Another technique for detecting an explicit item in a single image is referred as detecting on object. Various items in an individual image, such as a football game, an offensive player, a defensive player, a ball, and so on, are distinguished using enhanced object detection. To design a bounding box and classify everything inside it, an X, Y coordinate is used [8].

Edge detection is an image processing technique for identifying the pixel points in an image. [8].

Detecting the pattern another phase in finding the repetition found in colors, forms, and other visual markers present in the images is pattern detection [8].

Classifying the images is a technique that assembles the images into various categories [8].

Feature engineering is again one type of pattern detection which matches resemblances found in the image and to aid them to classify it [8].

II. **Machine Learning**

Machine can itself cannot search for patterns in the data. Hence the data to be fed into the computer should be pre-processed and transformed in to the form it recognizes. For such processes, one can utilize certain steps like classification, clustering, regression algorithms. These procedures depend on the information existing about the problem so that it could be ended with valuable results:

Classification: The algorithm that allocated labels to data based on the predefined features is called as classification. It is one of the process found in supervised learning. This algorithm sorts the data according to the patterns found in the images [9].

Clustering: This is an algorithm, which breakups data into quite number of cluster. It is based on the resemblance of features found in the images. This is also called as an unsupervised learning [9].

Regression: Relationship between different variables is found out by the Regression algorithms. It forecast the unknown dependent variables from the known data already available in the literature. It is again in an example of supervised learning [9].

III. **Artificial Intelligence (AI) in Image Processing**

The machines are trained to infer images in the same way as our brain works. In one step ahead, these machines can even examine the images much more systematically than it is possible for us. This methodology is applied in image processing, and popularly known as artificial intelligence (AI). This tool can investigate the process of face recognition and verifies the functionality for confirming the security in public places. It also recognizes and detects the pattern in images, videos, objects, and so on.

5 Classification of Plant Disease Using Classifiers

The major classifiers used in the recognition of plant diseases are Artificial Neural Networks (ANN) Convolutional neural network (CNN), K-Nearest Neighbors (KNN), Support Vector Machine (SVM) which are discussed below:

Convolutional neural network

A convolutional neural network is a specific type of neural network with several layers. The significant features in an image are extracted by means of processing the data using a grid-like arrangement. A CNN is network utilized to examine processing data in a grid-like architecture to create visual images [10]. Other names of CNN are feed-forward neural network, named as ConvNet. To identify and classifying the object in an image is usually done by a convolutional neural network. A major advantage of the CNN techniques is the steps involved in preprocessing on images are much reduced.

The key purpose of the convolutional neural network algorithm's is to process the data in simple way. Again, it does not lose any feature in the image which is vital for figuring out the representation of the given data. This truly has made them a promising candidate for managing huge datasets.

Furthermore, multiple hidden layers are present in the CNN which helps them in bringing out the details from an image. The 4 significant layers in CNN are:

1. **Convolutional Layer**: The convolutional layer is the initial phase in the CNN process, which involves extracting significant characteristics from an image [10]. This layer utilizes many filters that executes the convolutional operation. In this process, each image is taken into account as a matrix of pixel values.
2. **ReLu layer**: The rectified layer linear unit is ReLu. On achieving the extracted feature maps, the next process is taking them into the ReLu layer [10]. This layer performs an element-by-element procedure, converting all negative pixels

to 0. It causes nonlinearity in the network, and the resulting output is a rectified feature map.

3. **Fully Connected Layer**: Flattening is the fourth procedure, which is used to combine all of the 2-D arrays obtained from pooled feature maps into a single long continuous linear vector. When the flattened matrix was provided as input to the fully linked layer, the image was categorized [10].

Support vector machine

The hyperplane in an N-dimensional space is found and execute using the SVM algorithm. This N that denotes the various features definitely categorizes the data points. The important characteristics of this algorithm is that any possible number of hyperplanes can be selected to split up the two classes of data points. The maximum distance between and data points of two classes is found out using the hyperplane that defines the maximum margin.

K-nearest neighbors

Amidst supervised machine learning algorithms is the KNN algorithm. This could be utilized for two purposes namely, classification and regression in the analysis of the problems. In particular, the aim of KNN algorithm is to classify the predictive problems in industry. This KNN has two well defined properties namely:

Lazy learning algorithm

Since no specific training phase is available in KNN, it is simply called as a lazy learning algorithm. Again, it makes use of the complete data for training during classification.

Non-Parametric learning algorithm

The underlying data is not adopted in KNN and so it is also called as a non-parametric learning algorithm.

6 Conclusion and Research Direction

In this paper, generalized deep learning methods are conversed. Every concept is based on various techniques utilized for the identification of diseases in plants. Detecting diseases in plants at initial stage is imperative since it reduces the production of crops and generates economic crisis. Among different image processing approaches deep learning methodology supports in spotting the diseases automatically. But in manual finding it is not a simple task as it needs lot of effort, time, and money. The above survey delivers a right knowledge in the identification of disease in plants. Thus, researches can implement these approaches in the field of agriculture to identify the damages caused in the plants and solve the problem at early stage. This paper has revealed the several approaches in image processing technology utilized by researchers in the production of crops. The major steps involved in the process are feature extraction, segmentation and classification. The above study will provide the key sources of recognition and rectification of diseased plants at the primary stage. These methodologies will help the farmers to overcome the shortfall in the crop production. Various collections of images were utilized for determine the plant disease, the collection of images had healthy and diseased leaf images. Only minimum papers from the pool of plant disease research papers are examined in this review (Table 1).

This survey paper delivers a right knowledge in the identification of diseases in plants. These ideas and methodologies is possible to use it in the agriculture to identify the damages caused in the plants at earlier stages.

Table 1 Review on plant disease classification algorithm

Referred authors	Crop	Disease	Dataset	Method	Accuracy
Basavaraj et al. [11]	Paddy crop	Biotic and Abiotic Stressed	Original field image	Deep learning-based classification VGG-16 BPNN	92.89% 89.12%
Radhakrishnan et al. [12]	Rice	Rice Blast disease	Plant village	SVM	Computed from confusionMatrix
Ferentinos [2]	25 different crop	58 different classes of disease	Open Dataset of 87,848 images	Various deep learning models used	Accuracy is in Highest Success rate
Shanwen et al. [13]	Apple Cucumber	Alternaria, Mosaic, Rust Anthracnose, Angular leaf spot, Powdery Mildew	Not available	Fivefold Validation	90.43% 92.15%
Sue Han et al. [14]	14 crops	26 crop disease categories	Plant village	VGG-16, Inception V3, GoogLeNet	Accuracy is high
Lu et al. [4]	Rice			CNN	95.48%
Junde et al. [10]	Rice Maize	Leaf smut, leaf scald, Bacterial blight, White tip Gray leaf spot, Common Rust, Eyespot	Fujian Institute–china	VGGNet-19, DenseNet, Inception V3, ResNet 50	91.83%

(continued)

Table 1 (continued)

Referred authors	Crop	Disease	Dataset	Method	Accuracy
Zhang et al. [15]	Maize	Dwarf mosaic, leaf gray spot, northern leaf blight, brown spot disease	Plant village dataset	Googenet and CIFAR 10	GoogleNet-98.8% CIFAR 10–98
Raza et al. [16]	Tomato species	Powdery mildew fungus	A new feature set from the image data using global and local statistics	SVM with Linear kernel	90%
Picon et al. [17]	Wheat	Fusarium	Uploaded images from real field	CNN	96%

References

1. Barbedo JGA (2018) Factors influencing the use of deep learning for plant disease recognition. Biosystems Eng 172:84–91
2. Ferentinos KP (2018) Deep learning models for plant disease detection and diagnosis. Computers Elect Agriculture 145:311–318
3. Lu J, et al (2017) An in-field automatic wheat disease diagnosis system. Computers Elect Agriculture 142:369–379
4. Lu Y, et al (2017) Identification of rice diseases using deep convolutional neural networks. Neurocomputing 267:378–384
5. Brahimi M, Boukhalfa K, Moussaoui A (2017) Deep learning for tomato diseases: classification and symptoms visualization. Appl Artif Intell 31(4):299–315
6. Ramesh S, Vydeki D (2020) Recognition and classification of paddy leaf diseases using Optimized Deep Neural network with Jaya algorithm. Info Processing Agriculture 7(2):249–260
7. Mohanty SP, Hughes DP, Salathé M (2016) Using deep learning for image-based plant disease detection. Front Plant Sci 7:1419
8. Iqbal Z, et al (2018) An automated detection and classification of citrus plant diseases using image processing techniques: A review. Computers Electronics Agriculture 153:12–32
9. Sujatha R et al (2021) Performance of deep learning vs machine learning in plant leaf disease detection. Microprocess Microsyst 80:103615
10. Chen J, et al (2020) Using deep transfer learning for image-based plant disease identification. Computers Elect Agriculture 173:105393
11. Anami BS, Malvade NN, Palaiah S (2020) Deep learning approach for recognition and classification of yield affecting paddy crop stresses using field images. Artif Intell Agriculture 4:12–20
12. Radhakrishnan S (2020) An improved machine learning algorithm for predicting blast disease in paddy crop. Materials Today: Proceedings 33:682–686
13. Zhang S, et al (2018) Plant diseased leaf segmentation and recognition by fusion of superpixel, K-means and PHOG. Optik 157:866–872
14. Lee SH, et al (2020) New perspectives on plant disease characterization based on deep learning. Computers Elect Agriculture 170:105220

15. Zhang X, et al (2018) Identification of maize leaf diseases using improved deep convolutional neural networks. IEEE Access 6:30370–30377
16. Raza S-A, et al (2015) Automatic detection of diseased tomato plants using thermal and stereo visible light images. PloS One 10(4):e0123262
17. Revathi P, Hemalatha M (2012) Classification of cotton leaf spot diseases using image processing edge detection techniques. In: 2012 International Conference on Emerging Trends in Science, Engineering and Technology (INCOSET). IEEE

Extreme Gradient Boosting for Toxic Comment Classification

Mehul Jain, Manas Suryabhan Patil, and Chandra Mohan

Abstract Increasingly, online discussions have been subject to offensive comments and abuse from malintent users. The threat of abuse and online harassment often forces many people to withdraw from discussions or shut down essential conversations. The current explosion in the number of online communities poses a huge problem for online social media platforms in terms of effective moderation of discussions and prevention from abuse. This paper aims to present a scalable method for effectively classifying toxic comments. Recently, the XGBoost framework has become very popular among machine learning practitioners and data scientists due to its effectiveness, flexibility, and portability. It exploits the gradient boosting framework and adds several improvements that enhance the performance and accuracy of the model. Due to its desirability, we utilize XGBoost to create an efficient system for toxic comment classification and evaluate its performance using area under curve and receiver operating characteristic curve.

Keywords XGBoost · Ensemble learning · Natural language processing · Machine learning

1 Introduction

Social media has given us the ability to communicate with each other quickly, effectively, and effortlessly. With the help of social media platforms, people from all around the globe can contribute to discussions regarding important issues and

M. Jain (✉) · M. S. Patil · C. Mohan
School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India
e-mail: mehul.jain2018@vitstudent.ac.in

M. S. Patil
e-mail: manassuryabhan.patil2018@vitstudent.ac.in

C. Mohan
e-mail: chandramohan.b@vit.ac.in

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_29

share their perspectives. For the most part, this enables people from different backgrounds to communicate with each other, but occasionally, it also results in truculent arguments, unproductive debates, and online harassment.

A significant problem for large social media platforms like Facebook, Twitter, Reddit is to mitigate the damage caused by offensive and violent comments. Violent comments often shut down meaningful conversations and discourage users from expressing their opinions or seeking different views. This impairs the free exchange of ideas and ends up harming the online ecosystem. In this paper, we demonstrate the utility of natural language processing and machine learning in identifying offensive comments.

The previous works have explored the usage of both conventional machine learning models such as support vector machines (SVM), logistic regression, and deep learning models such as convolution neural network (CNN). This work explores the usage of XGBoost in making powerful models for detecting toxicity in comments. XGBoost has recently exploded in popularity among data scientists and machine learning practitioners alike. The XGBoost framework is known for its scalability, flexibility, and portability. Using XGBoost, we build a binary classifier to differentiate between toxic and non-toxic comments. The system proposed is evaluated using area under curve (AUC) and receiver operating characteristic curve (ROC). These metrics permit us to perform a detailed evaluation of the model performance.

2 Literature Survey

The construction of an efficient machine learning model requires a sufficient amount of high-quality labeled data. A comprehensive methodology for generating large-scale longitudinal data on personal attacks in online discussions is presented in [1]. Furthermore, this study provides a framework for building a binary classifier that can be used for analyzing online toxicity. The dataset used in the paper was made publicly available for use on Kaggle in a competition named 'Toxic Comment Classification Challenge.' The challenges in detecting unintended bias present in the toxicity classifier [2] are described in literature. Furthermore, it also presents novel approaches and countermeasures to mitigate bias present in the machine learning model.

Many works [3–5] in the field propose deep learning techniques to perform toxic comment classification. Although deep learning approaches have some significant advantages, the focus of this paper is to present a scalable approach that can be employed on larger datasets and performs better in comparison to other feasible approaches. The author [6] proposes a multi-class classification of toxic comments, identifying subclasses like severely toxic, obscene, threat, insult, etc. Furthermore, the paper also explores the lack of sufficient data for each subclass and propounds the idea of contextual word embedding to increase the accuracy of the classifier.

The authors [7] delineate all the challenges and possible caveats for toxic comment classification systems. The study provides an in-depth error analysis which is beneficial in understanding the model predictions and evaluating false positives and false negatives. It also provides a detailed comparison of different model architectures tested on two datasets using different language models (word embedding, character n-grams, word n-grams) to represent the training data. The paper [8] claims that text preprocessing improves model accuracy which is questioned and evaluated with multiple preprocessing techniques. Based on empirical evidence, the authors recommend focusing on model selection instead of spending much time on the text preprocessing steps.

3 Proposed Methodology

In this section, the methodology of the proposed approach is presented. Figure 1 shows the framework of our proposed method.

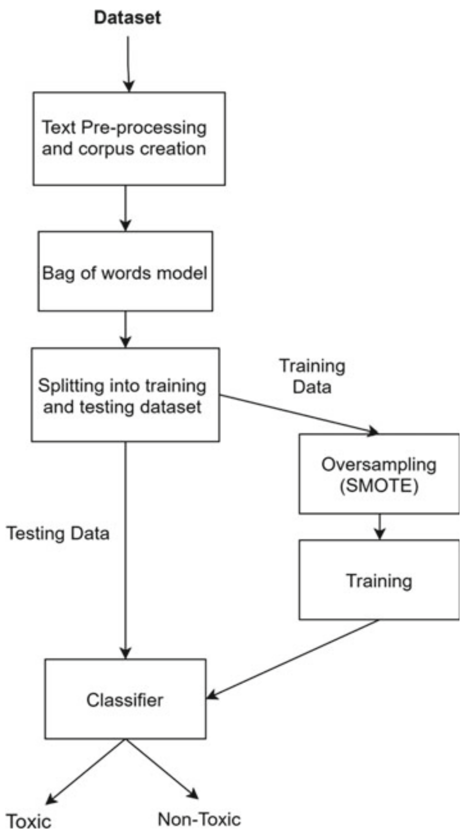


Fig. 1 The framework of the proposed approach

3.1 Corpus Creation and Text Preprocessing

The initial step in creating the toxic comment classification system is corpus creation and text preprocessing. A corpus refers to a collection of text documents that are used to train the machine learning model. Each comment in the dataset is first preprocessed and then added to the corpus. The following steps are applied for preprocessing the comments in our dataset:

- **Removing hyperlinks**: Often, text documents contain hyperlinks that point to references or other resources that might be useful for a reader. While creating a text classification model, hyperlinks need to be removed since they do not contribute any valuable information for the categorization of a text document.
- **Removing punctuation marks and numbers**: Text documents contain a variety of punctuation marks and numbers in different parts of sentences. They help clarify the meaning of a sentence, but they are not relevant to the overall context of the data and the final output from the text classification model. Removing punctuation marks simplifies the text documents, which in turn facilitates the model building process.
- **Converting text documents to lower case**: Converting text documents to lower case ensures that the model does not treat a word with capital letters differently from the same word, which appears later in the sentence but without any capital letters.
- **Removing stop words from the text**: Stop words are the words in a language that do not contribute significantly to the meaning of a sentence. In English, these are words like a the, an this, such, etc. They are removed from the text for numerous reasons. Since stop words occur in abundance, they provide negligible or no unique information that can be used for classification. On removing stop words, the corpus size decreases; this simplifies the model and reduces the time taken for training. Finally, stop word removal also improves accuracy since only the tokens with significant weightage remain.
- **Stemming and lemmatization**: Often, documents use different forms of a word to support different grammatical contexts, such as run, ran, and running. In addition, there are families of words that are derivationally related, for instance, merit, merited, and meritocracy. In such situations, stemming and lemmatization are used to reduce derived words or inflectional forms of words to a common base form. Stemming relates to the process of removing the ends of words or derivational affixes with the aim of converting words to a common base form. On the other hand, lemmatization points to the use of vocabulary and morphological analysis of words, intending to remove inflectional endings so as to return the base or dictionary form of a word, which is called the lemma.

3.2 Corpus Sampling and Bag-of-Words Model

After text preprocessing and corpus creation, we move forward towards creating a bag-of-words model to represent the textual data for modeling. In a bag-of-words (BoW) model, a text document is described by the number of occurrences of each word in it. In our case, each row in the BoW represents a comment, while each column represents a word from the vocabulary. Each cell in a given row represents the number of times a word occurs in a given comment.

One of the key challenges with a BoW model is the high amount of sparsity present when working with a large corpus that utilizes a vast vocabulary. Due to the high sparsity and limited computational resources, the corpus needs to be downsized before the creation of a BoW model. In order to accomplish this, we employ random sampling. In this process, appropriate sample size is chosen, and samples are randomly picked from the collection of documents until the desired sample size is attained. The sample size is heavily dependent on the computational resources available and the size of accessible memory. The resulting corpus after random sampling is used to create the BoW model.

3.3 Oversampling

Figure 2 depicts the number of instances in the positive (toxic) and negative (non-toxic) classes in the dataset. We can observe that the dataset is highly imbalanced; it contains a disproportionately high number of negative cases (non-toxic) and a lesser number of positive cases (toxic). An imbalanced dataset can be detrimental to the training process and may result in the creation of a model that only performs well for detecting instances of the majority class. Oversampling is the process of generating or duplicating samples in the under-represented classes in order to compensate for the imbalance in the dataset.

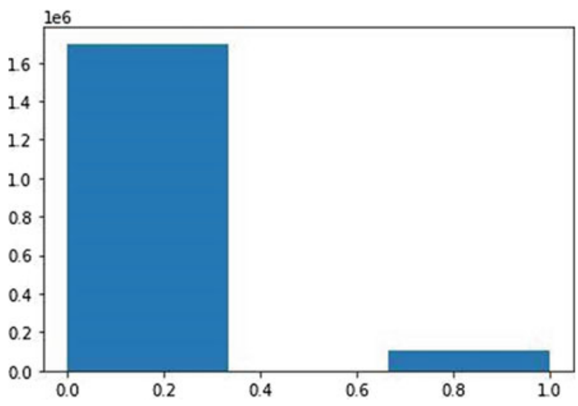


Fig. 2 The balance between the positive and negative classes in the original dataset

Synthetic minority oversampling technique (SMOTE) is an oversampling technique that is used to synthesize new instances from the existing instances in order to treat the imbalance. From the paper [9], SMOTE operates by selecting instances that are closely located in the feature space and generating a new instance on the line joining the selected instances. The newly generated samples are comparatively closely located to the existing samples in the feature space. As a result of this, SMOTE proves to be effective in compensating for the imbalance present in the dataset. Furthermore, it is also computationally efficient in comparison with more sophisticated data augmentation techniques. The re-sampled BoW formed after using SMOTE is used as the training set in the succeeding stages for model training and evaluation.

3.4 Gradient Boosting and XGBoost

In this step, we use the re-sampled BoW created in the previous step and build our binary classifier. As mentioned earlier, we use the XGBoost framework to create our classifier. XGBoost stands for extreme gradient boosting. From [10], XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. Boosting is an ensemble learning technique that aims to create a robust model using several weaker models in a sequence.

Initially, a model is built from the original training set. Then, a second model is built that aims to rectify the errors from the previous models. This process continues until either the training set is fit perfectly or the maximum number of models specified is added. Gradient boosting reshapes the technique of boosting and converts it into a numerical optimization problem. In gradient boosting, the goal is to minimize the loss by adding weak learners using the process of gradient descent. Gradient descent is an iterative process for finding the local or global minimum of a given differentiable function. In the beginning, a model is created to predict the values of the target variables. Then, iteratively, models are created that predict the residual errors from the previous models. In this fashion, a series of models are created such that the contribution of each model minimizes the overall error of the ensemble. This can be represented mathematically.

Let us consider a gradient boosting scenario with N stages. At any given stage n (n lies between *1* and N both inclusive), an imperfect model F_n may return $\widehat{y_i}$. In order to improve F_n, the algorithm should add another predictor $h_n(x)$, refer to Eq. (1).

$$F_{n+1}(x) = F_n(x) + h_n(x) = y \tag{1}$$

$$h_n(x) = y - F_n(x) \tag{2}$$

From Eq. (2), the algorithm will fit the new predictor *h* to the residual. Each subsequent F_{n+1} model aims to rectify the errors of the previous models. In the end, the final prediction is the sum of the predictions of all the base learners.

XGBoost is an extension of the GBMs (gradient boosted decision trees), which is specifically devised to give better performance and train in less time. The XGBoost framework uses the following techniques to create an accurate model and improve on the preexisting gradient boosting framework:

- **Shrinkage and column sub-sampling**: In XGBoost, shrinkage is employed for lessening the influence of each additionally fitted weak base model. It penalizes the impact of each successive iteration by reducing the size of incremental steps. This way, the negative impact of an erroneous iteration can be compensated for by the successive steps. Another technique used by XGBoost to avoid overfitting is column sub-sampling. As in random forest, a proportion of columns is randomly selected from the feature set to fit a model. This process introduces randomness to the training process and improves generalization.
- **Gradient Boosting**: The ensemble model is optimized using a sequential and iterative process, where decision trees are added in each iteration. The loss value reduces with each iteration, and the overall error decreases. As mentioned earlier, gradient boosting is stopped either when the specified number of trees are added to the model or the loss reaches a tolerable value.
- **Regularization**: Regularization penalizes the weights in the final iterations so that the model does not over-fit. It also gravitates towards choosing models which apply straightforward and predictive functions. Our model takes in many hyperparameters, which can be optimized to give a better result. Some of the hyperparameters include the number of trees, learning rate, maximum depth of a tree, regularization parameter (gamma) value, etc. Here, we apply the random search optimization technique to find the ideal values for the hyperparameters. Other techniques such as grid search and Bayesian search can also be applied based on the availability of computational resources.

3.5 Model Evaluation

The XGBoost model created is trained on the re-sampled training set and evaluated on the validation set. To evaluate the predictions generated by our XGBoost model, we use the ROC curve and AUC. The ROC curve depends on the following two metrics:

- **True Positive Rate/ Recall**: Recall expresses the proportion of the positive class that was correctly classified by the model. Refer to Eq. (3).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

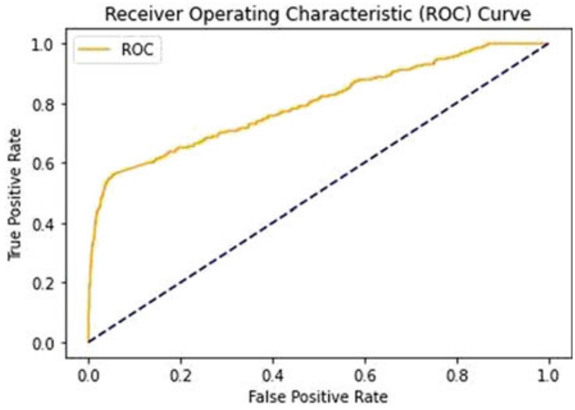


Fig. 3 No. of training samples: 20,000, no. of features:10,000

- **False Positive Rate:** FPR expresses the proportion of the negative class that was incorrectly classified by the model. Refer the Eq. (4).

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}} \tag{4}$$

Using the above metrics, the ROC curve and AUC are defined as:

- **ROC curve**: ROC is a probability plot that graphs the True Positive Rate against the False Positive Rate at various threshold values for classification. Figure 3 is an example of a ROC curve.
- **AUC**: AUC reports the ability of the model to differentiate between the positive and the negative class. Another way of interpreting AUC is as the probability that the model ranks a random positive. In general, the higher is AUC, the better the model is at differentiating the instances of the positive and the negative class.

4 Results and Discussion

In the previous section, we presented our proposed approach and explained the various steps in detail. In this section, we present our experimental setup and the results obtained.

In corpus creation and text preprocessing, the details for creating a corpus are presented. Using the corpus, we create a BoW model to represent our textual data. Since the computational resources are limited, we need to specify a maximum number of features (N) that should be used in the BoW. Based on this number, a vocabulary is built that only contains the top N features (words) ordered on the basis of frequency of occurrence in the corpus. Table 1 shows the number of comments and the number of maximum features used to build the BoW model in each iteration of the experiment.

Table 1 Configuration for BoW in each iteration and the AUC score achieved

Serial No	No. of comments (samples)	No. of features	AUC
1	20,000	10,000	0.80
2	20,000	20,000	0.85
3	20,000	30,000	0.86
4	30,000	10,000	0.86
5	30,000	15,000	0.88

For instance, 20,000 comments and 10,000 features (words) were used to build a BoW model in the first iteration.

Hyperparameter tuning is an essential step in ensuring that our model uses the optimal values for the hyperparameters. Sub-optimal values can result in a model that has lower accuracy and does not generalize well on unseen data. As mentioned in gradient boosting and XGBoost, we use the random search optimization technique to obtain the ideal values for the hyperparameters used in the XGBoost model. Applying random search gives us the following values for the hyperparameters of our XGBoost model:

- Number of gradient boosted trees: 300
- Maximum depth of trees: 5
- Learning rate: 0.1
- Threshold score: 0.5

It is important to note that these values can differ with the size of the dataset and the number of features. They do not represent the optimal values in all scenarios, but in our experiments, they appear to be the optimal values.

We evaluate our toxic comment classifier using the ROC curve and AUC. Table 1 shows the BoW model created in each iteration, along with the corresponding AUC score. Figures 3, 4, 5, 6, and 7 show the ROC curve for each row of Table 1. As

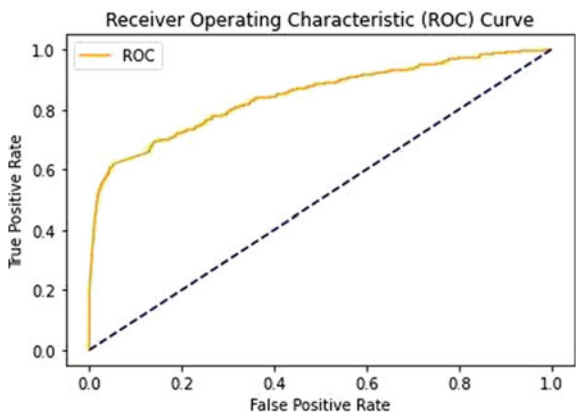


Fig. 4 No. of training samples: 20,000, no. of features: 20,000

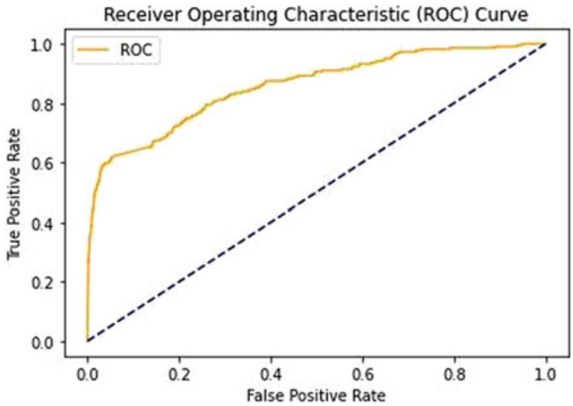


Fig. 5 No. of training samples: 20,000, no. of features: 30,000

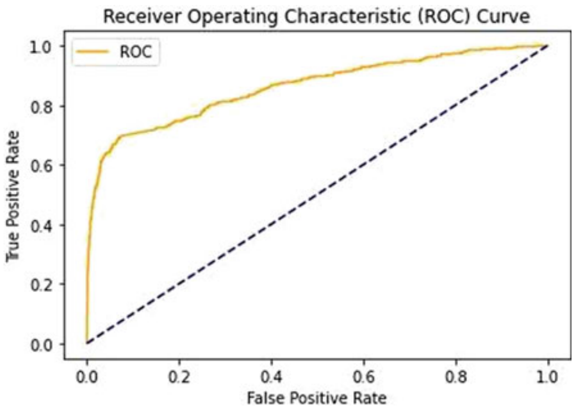


Fig. 6 No. of training samples: 30,000, no. of features: 10,000

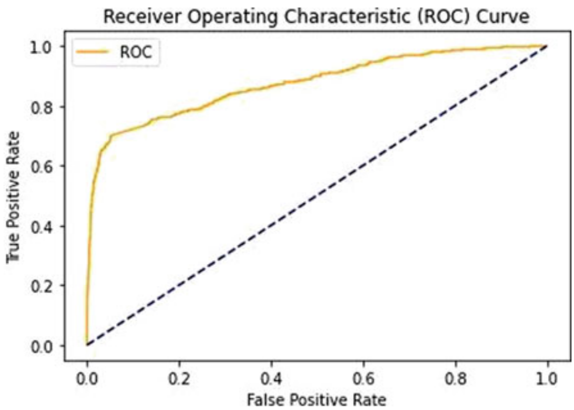


Fig. 7 No. of training samples: 30,000, no. of features: 15,000

you can observe, increasing the size of the training set by adding more comments results in better performance by the toxic comment classifier. Furthermore, we also notice that using a larger vocabulary results in a dramatic improvement in the model performance.

5 Future Work

There are several possible directions for future work. These include ensembling different custom models to create a classification system that is more accurate and presents a more realistic picture of online toxicity. Additionally, with the advancements in the field of language modeling and the increasing applications of word embeddings, we hope to substitute the bag-of-words model used in this work with a pre-trained word embeddings model. Furthermore, there is scope to improve the model robustness using data augmentation and adversarial training approaches. Finally, we hope to introduce alternative customized metrics that help in understanding the model fairness and also indicate the pitfalls of the generated model.

6 Conclusions

In this paper, we present a methodology to build a scalable and accurate binary classifier to differentiate toxic and non-toxic (normal) comments. We demonstrated how the XGBoost framework can be applied to textual data to perform complex text classification tasks. Furthermore, we also showed the effect of different parameters on model performance using the ROC curve and AUC. The methodology introduced in this paper has wide-ranging applications such as flagging abusive users, detecting toxic comments automatically, and allowing for preemptive moderation

In the course of our research, we realized that the majority of online comments are not hurtful, but a minority of comments that are offensive end up demotivating users and decreasing user participation. The ever-growing accessibility to the World Wide Web and the increasing number of large online communities has allowed us to communicate with people and groups which otherwise were inaccessible. Preserving the sanctity of these conversations is an important issue we face going forward, and it demands novel approaches which prevent conflicts and promote communication.

References

1. Wulczyn E, Thain N, Dixon L, Machina E (2017) Personal attacks seen at scale. In: Proceedings of the 26th International Conference on World Wide Web, April. https://doi.org/10.1145/3038912.3052591
2. Dixon L, Li J, Sorensen J, Thain N, Vasserman L (2018) Measuring and mitigating unintended bias in text classification. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, Feb. https://doi.org/10.1145/3278721.3278729
3. Georgakopoulos SV, Tasoulis SK, Vrahatis AG, Plagianakos VP (2018) Convolutional neural networks for toxic comment classification. In: Proceedings of the 10th Hellenic Conference on Artificial Intelligence—SETN '18. https://doi.org/10.1145/3200947.3208069
4. Carta S, Corriga A, Mulas R, Recupero DR, Saia, R (2019) A supervised multi-class multi-label word embeddings approach for toxic comment classification. KDIR, pp 105–112, Sep
5. Chu T, Jue K, Wang M (2021) Comment abuse classification with deep learning. Accessed Aug 20. Available online at:https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2762092.pdf
6. Pavlopoulos J, Malakasiotis P, Androutsopoulos I (2017) Deep learning for user comment moderation, July [Online]. Available: https://arxiv.org/abs/1705.09993
7. Van Aken B, Risch J, Krestel R, Loser A (2018) Challenges for toxic comment classification, Sep [Online]. Available: https://arxiv.org/abs/1809.07572
8. Mohammad F (2018) Is preprocessing of text really worth your time for online comment classification? Aug. Accessed Aug 20 [Online]. Available: https://arxiv.org/abs/1806.02908
9. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: Synthetic minority oversampling technique. J Artificial Intelligence Res 16:321–357. https://doi.org/10.1613/jair.953
10. Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining—KDD '16. https://doi.org/10.1145/2939672.2939785

Machine Learning Models for Predicting Customer Willingness to Buy Electric Vehicles

R. Madhu Shri and S. N. Vivek Raj

Abstract Countries across the world have started taking measures to control vehicular pollution, and one great initiative is to implement the Electric Vehicles. In India, where vehicular density is high, it is important to implement Electric Vehicles at a faster pace. Every innovation lacks promotion which will aid its success. This study predicts the customer willingness to purchase an electric vehicle and analyses the significant predictors of the Electric Vehicle purchase willingness. A sample of 371 respondents are considered, and convenience sampling is used to collect data over the Internet. Machine learning models are built with the available data and of all the machine learning algorithms, binary logistic regression, random forest and gradient boosting have the highest accuracy. Prior Experience, Cost Convenience, Aesthetic Convenience, Conventional Vehicle usage, Individual attitude and Gender are the most significant predictors of the purchase willingness. Also, through the sentiment analysis, it is known that people largely consider electric vehicles as good and eco-friendly.

Keywords Electric vehicles · Machine learning · Logistic regression · Decision tree · Sentiment analysis

1 Introduction

Electric Vehicles are vehicles that use electric motors in place of an Internal Combustion Engine. Electric vehicles are slowly entering into the market, spreading wings as an alternate to conventional vehicles. Electric Vehicles are encouraged because they are environment friendly. All over the world, the governments and Electric

R. M. Shri · S. N. V. Raj (✉)
KCT Business School, Kumaraguru College of Technology, Coimbatore, India
e-mail: vivek@kctbs.ac.in

R. M. Shri
e-mail: madhushri.20mba@kct.ac.in

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_30

Vehicle sellers are framing various strategies to make people prefer Electric Vehicles. The study puts forth prediction models, to predict whether a customer is willing to purchase an Electric Vehicle or not, based on certain characteristics. This prediction paves an easy way for the sellers to segment right customers through whom they can promote the sales of Electric Vehicles. The significant predictors of the customer willingness are also understood from the study. Once these customers buy, this can create awareness among the people around. This study serves as an initial spark by bringing in machine learning to set up potential customers and understand overall perception of the customers to buy electric vehicles. The primary objective of the study is to predict customer willingness to buy electric vehicles by developing machine learning models and assessing them. Other than this, a text analysis using word cloud is made to visualize the prevalent subjective opinions of the customers. Through this, several ideas can be suggested to government and sellers to make people adopt electric vehicles. The paper has 7 sections apart from the Introduction section. In Sect. 2, the literature review is presented, in Sect. 3, the research methodology that includes the research design, sampling design and the hypotheses are discussed, in Sect. 4 the data is analysed with the machine learning algorithms and sentiment analysis, in the Sect. 5, findings and suggestions are listed. The scope and limitations of the research are discussed in Sect. 6, and the study is concluded in Sect. 7. Further, the references are listed in Sect. 8.

2 Review of Literature

Several studies were made on analysing the customer willingness and behaviour towards buying an Electric Vehicle. Structured Equation Modelling [1, 5, 7–9, 11, 15, 16] is the widely used technique in most of the studies to analyse the relationship among the constructs. Other than that regression [1, 14, 16], factor analysis [1, 3] and artificial neural networks [16] were also used in determining the relationship between various constructs and the purchase intention. While selecting the constructs, models such as TPB (Theory of Planned Behaviour), TRA (Theory of Reasoned Action) and TAM (Technology Assessment Model) were used [1, 8, 10, 14–16]. The constructs considered include Demographic factors, Environmental consequences, cost and convenience, individual consequences, values, consumer trust, comfort, technology, infrastructure, social acceptance, subjective norm control, perceived behaviour, psychological and policy attributes. Apart from these, price value, non-monetary and monetary incentive policies, operation cost, economic evaluation, brand identity, brand awareness, perceived risk, pro environmental behaviour and performance features were also considered. Different tools used for Data collection, sample size determination and analysis are SPSS [2, 6, 14, 16], G*power software [14], fsQCA [17] and SMART PLS [16]. Most of the studies concluded that Environmental consequences as one of the significant predictors of the purchase intention. Cost, infrastructure, opinions of friends, brand awareness, financial advantages and social acceptance did not have an impact on the customer willingness [3, 6]. Perceived behaviour, monetary policies, subjective norm, attitude and brand image are strongest predictors of purchase intention. Other inferences made from the studies indicate that middle age, professional and rural/sub-urban multi-person household [12] are very likely to buy Electric Vehicles and the adoption of Electric Vehicles is possible when proper awareness is given [14]. The research studies widely implemented Structural Equation Modelling (SEM) and factor analysis to understand purchase intention, but no study was done on the prediction of customer willingness to buy Electric Vehicles. The aesthetic properties and operational convenience of electric vehicles were not largely considered in the earlier studies. The cost, personal attitude and environmental awareness were the most prioritized in determining a customer's purchase intention. There were no subjective responses considered. Some studies involved collecting demographic factors and some variables based on 5-point [14] and 7-point scale [1, 5, 7]. Adding to this, the research studies suggested promotional strategies based on the variables analysed and their significance in purchase intention did not cover specific customer-targeted promotion.

3 Research Methodology

3.1 Research Design

The type of research is predictive research using CRISP-DM methodology, and the study involves collection of primary data through a structured questionnaire in Google Forms and is circulated to the respondents over the Internet due to the time constraints and the COVID-19 pandemic. The sampling technique used for the research is Convenience sampling. After cleaning the data, 371 responses are selected and this is sufficient as per the method mentioned by R. Afroz, A. Rahman, M. M. Masud, R. Akhtar and J. B. Duasa, where sample size used should be more than 15 times the number of constructs [1]. The variables considered are based on the extended TPB (Theory of Planned Behaviour Model) [1] that includes Perceived Behaviour, Subjective Norm, Attitude, Behavioural intention. Along with this, other variables are considered to increase the scope of the research and hence 12 hypotheses were framed. The model studies the behaviour of a person based on their own beliefs. The willingness does not just depend on personal beliefs but other factors like the demographic factors (gender, age, family size, family annual income), the type of vehicle usage, prior experience with a particular vehicle, distance travelled daily and the kind of vehicles used by friends can also have unnoticed effects on the purchase willingness. The machine learning models used can take in continuous, discrete and categorical variables as independent variables. Hence, all possible factors are considered. The constructs Individual Attitude, Environmental Awareness, Subjective Norm, Awareness on Government Policy Measures, Operational Convenience, Aesthetic Convenience and Cost Convenience have 3–4 items under them on the five-point scale, and the average is considered for each.

3.2 Hypotheses

The hypotheses are as follows.

H1: Demographic factors are significant predictors of Electric Vehicle purchase willingness, **H2**: Conventional vehicle usage is a significant predictor of Electric Vehicle purchase willingness, **H3**: Travelling distance is a significant predictor of Electric Vehicle purchase willingness, **H4**: Friends Electric Vehicle usage is a significant predictor of Electric Vehicle purchase willingness, **H5**: Prior Experience with Electric Vehicle is a significant predictor of Electric Vehicle purchase willingness, **H6**: Individual attitude is a significant predictor of Electric Vehicle purchase willingness [1], **H7**: Environmental awareness is a significant predictor of Electric Vehicle purchase willingness [9], **H8**: Awareness on Government policies is a significant predictor of Electric Vehicle purchase willingness [16], **H9**: Subjective norm is a significant predictor of Electric Vehicle purchase willingness [15, 16], **H10**: Aesthetic convenience is a significant predictor of Electric Vehicle purchase willingness [1], **H11**: Cost convenience is a significant predictor of Electric Vehicle purchase willingness [1, 5], **H12**: Operational convenience is a significant predictor of Electric Vehicle purchase willingness.

4 Data Analysis

4.1 Reliability Test and Factor Analysis

See Table 1.

After collection of the responses, the data was checked for reliability by taking around fifty samples. The reliability among the questions under each construct was checked before starting with the machine learning algorithms. The test is done in SPSS. Cronbach alpha and Spearman–Brown coefficients are checked for each construct. All values are above 0.7 and showed a good reliability, the ideal values of Cronbach alpha and Spearman–Brown coefficient being more than 0.6 [16]. These values represent good internal consistency of the constructs. Composite reliability is expected to be above 0.6, and the Average Variance Extracted should be above 0.50 [16]. The values met the criteria, thus ensuring construct validity.

4.2 Machine Learning Algorithms

Binary logistic regression and decision tree are the algorithms used to predict the willingness and analyse the predictor variables in SPSS. Apart from these, algorithms such as KNN, Random Forest and boosting algorithms such as AdaBoost

Table 1 Reliability test and factor analysis

Constructs	Cronbach alpha	Spearman–Brown coefficient	Factor loadings	Composite reliability	Average variance extracted
Individual attitude (IA)	0.906	0.904	0.909–0.923	0.848	0.954
Environmental awareness (EA)	0.921	0.877	0.898–0.945	0.864	0.95
Awareness on government policy measures (GPM)	0.939	0.885	0.869–0.928	0.854	0.959
Subjective norm (SN)	0.815	0.804	0.848–0.859	0.731	0.891
Aesthetic convenience (AC)	0.948	0.949	0.945–0.958	0.906	0.967
Cost convenience (CC)	0.896	0.923	0.886–0.941	0.827	0.935
Operational Convenience (OC)	0.883	0.742	0.777–0.967	0.929	0.815
Purchase intention (PI)	0.983	0.99	0.967–0.992	0.968	0.989

and Gradient Boosting are implemented in Python to test model performances in predicting the willingness.

Binary Logistic Regression: Binary logistic regression is a classification algorithm used to predict binary output. The variable is coded with 0 (not willing) and 1 (willing). The analysis is done in SPSS. Forward and Backward methods are implemented in SPSS.

The categorical variables are coded as shown in Table 2. For 'n' classes of variables, 'n-1' dummy variables are considered. The columns show the dummy variable labels and '1' is given under the corresponding label for each category of the variable.

Binary Logistic Regression Forward

The forward binary logistic regression involves step-by-step prediction. The first step considers the most significant predictor. Then subsequently, the next highest significant predictors are considered. The significant predictors are the Cost Convenience, Prior Experience, Owning a conventional vehicle, gender, individual attitude and aesthetic convenience. The other variables are non-significant and are not considered. People with prior experience are 0.855 times more likely to buy than the people with no experience and people who do not own a conventional vehicle are more

Table 2 Categorical variable coding

		Parameter coding			
		(1)	(2)	(3)	(4)
Family annual income	10–20 lakh	1.000	0.000	0.000	0.000
	20–30 lakh	0.000	1.000	0.000	0.000
	5–10 lakh	0.000	0.000	1.000	0.000
	Below 5 lakh	0.000	0.000	0.000	1.000
	More than 5 lakh	0.000	0.000	0.000	0.000
Family size	2 and lesser	1.000	0.000	0.000	
	3	0.000	1.000	0.000	
	4	0.000	0.000	1.000	
	5 and above	0.000	0.000	0.000	
	Urban	0.000	0.000	0.000	
Prior experience	No	1.000			
	Yes	0.000			
Conventional vehicle usage	No	1.000			
	Yes	0.000			
Friends' electric vehicle usage	No	1.000			
	Yes	0.000			
Gender	Female	1.000			
	Male	0.000			

likely to buy. In case of gender, women are 0.487 times less likely to buy the Electric Vehicles (Table 3).

Binary Logistic Regression—Backward

In case of the backward propagation, the least significant variable is discarded step-by-step and in the final step, only the significant variables are considered. The

Table 3 Final step of forward logistic regression

		B	S.E	Wald	*df*	Sig.	Exp (B)
Step 6	Individual attitude	−0.407	0.166	5.981	1	0.014	0.666
	Aesthetic convenience	0.406	0.183	4.887	1	0.027	1.500
	Cost convenience	0.615	0.192	10.228	1	0.001	1.849
	Gender (1)	−0.487	0.239	4.140	1	0.042	0.614
	Conventional vehicle usage (1)	1.285	0.451	8.112	1	0.004	3.616
	Prior experience (1)	−0.855	0.256	11.165	1	0.001	0.425
	Constant	−0.607	0.502	1.463	1	0.226	0.545

Table 4 Final step of backward logistic regression

		B	S.E	Wald	df	Sig.	Exp (B)
Step 11	Individual attitude	−0.407	0.166	5.981	1	0.014	0.666
	Aesthetic convenience	0.406	0.183	4.887	1	0.027	1.5
	Cost convenience	0.615	0.192	10.23	1	0.001	1.849
	Gender (1)	−0.487	0.239	4.14	1	0.042	0.614
	Conventional vehicle usage (1)	1.285	0.451	8.112	1	0.004	3.616
	Prior experience (1)	−0.855	0.256	11.16	1	0.001	0.425
	Constant	-0.607	0.502	1.463	1	0.226	0.545

results are like the forward method, with the final significant predictors being—Individual Attitude, cost convenience, aesthetic convenience, gender, prior experience and conventional vehicle usage. Men, people with prior experience and people who do not own a conventional vehicle are more willing to buy an electric vehicle (Table 4).

Decision Tree

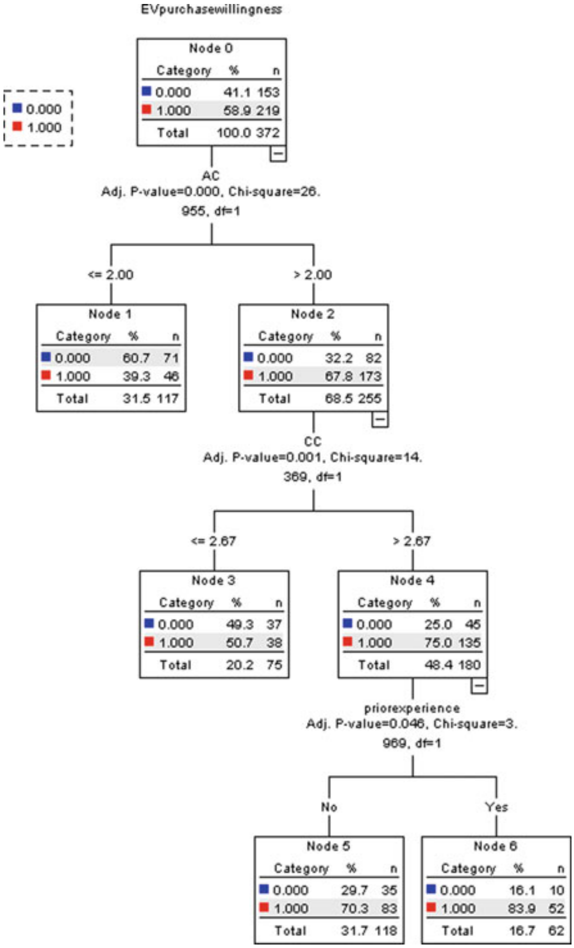
The decision tree uses tree like structure to predict outputs. It uses parameters such as entropy, Gini impurity and information gain to identify the root node and the following child nodes. The nodes represent features of the data, and the leaf nodes are the results. Here, CHAID algorithm is used to build the decision tree in SPSS.

In the output, for each node, percentage of people willing to buy and not willing to buy is given. As the decision tree root node is decided based on its importance and the first node here is aesthetic convenience, it is concluded that the aesthetic convenience has the highest importance from the decision tree model. The root node is classified into two child nodes—one with less than 2 and another with greater than 2 (which apparently is the average of the items under 'aesthetic convenience'). In a similar way, the next important characteristics include Cost convenience and prior experience in using electric vehicles. The terminal nodes (the nodes that are not classified further) are the ones from which we will be able to know the predicted value. Node 3 did not classify it with greater accuracy since it classified approximately half of population as 1 and the other half as 0. The terminal nodes 5 and 6 of the decision tree represent 'no' and 'yes' classes of prior experience, with 70.3% willing to buy without prior experience and 83.9% willing to buy with prior experience (Fig. 1).

4.3 Sentiment Analysis

A separate column for subjective response is recorded, and text analysis is done. The text analysis is done through generating a word cloud in Python and getting to know the most frequent words used by people to describe what they feel about electric vehicles in general. By looking at the word cloud, the most frequently used word is

Fig. 1 Decision tree



'electric vehicle' which is obvious. The next majorly used words as we can see are 'good', 'environment friendly', 'eco-friendly', 'future', 'petrol', 'cost', 'charging', 'reduce', 'fuel', 'will' and 'pollution'. And then there are other less significant words. From the frequently used words, we could interpret that most people think that electric vehicles are environment friendly and reduce pollution. 'Future' is the word and we can infer that electric vehicle penetration will increase in future since there is not much say about the present condition. On the other hand, it is also seen from the word cloud that people are concerned about the charging facilities and has clear understanding that electric vehicles can reduce fuel consumption and pollution. The most frequent word used next to 'electric vehicle' is 'good', from which we have a great inference that there is a positive perception towards electric vehicles among the people (Fig. 2).

Fig. 2 Word cloud

4.4 Performance Measures of Different Algorithms Used

Precision indicates the proportion of positive predictions that are actually correct [4, 13] and the ideal value should be 1 (when there are no false positives). Recall indicates the proportion of actual positives correctly classified and the ideal value is 1.0 (no false negatives) [4, 13]. F1-score [4, 13] is the combination of precision and recall. Binary logistic regression has the highest precision (0.68). Again, binary logistic regression has the highest F1-score (0.62) followed by Random Forest (0.61) and Gradient boosting (0.61). Binary Logistic Regression, Random Forest and Gradient Boosting have the highest accuracy and recall among all the algorithms (0.62) (Table 5).

Table 5 Performance measures

Algorithm	Precision	Recall	F1-score	Accuracy
Binary logistic regression	0.68	0.62	0.62	0.62
Decision tree	0.59	0.59	0.53	0.59
kNN	0.5	0.52	0.51	0.52
Random forest	0.61	0.62	0.61	0.62
AdaBoost	0.59	0.6	0.6	0.6
Gradient boosting	0.61	0.62	0.61	0.62

5 Findings and Suggestions

Among the algorithms that were implemented, the binary logistic regression, gradient boosting and random forest have the highest accuracy. Prior experience was a significant predictor under both binary logistic regression and decision tree. By analysing the log odds ratio, people who had prior experience in driving electric vehicles preferred electric vehicles the most. Through this, we get to know that there are higher chances for a customer to buy an electric vehicle if he/she has used it earlier and got to know the pros of the vehicle. The sellers can offer free test rides to the customers, say for a week or month so that they will get to know the pros and cons of electric vehicles. Free test rides can be started in colleges to increase awareness among students who can expedite the Electric Vehicle usage. They can also bring into market-rented electric vehicles like 'Bounce' vehicles in Bangalore, India, that would be very much helpful in people thinking of buying own electric vehicles. Cost convenience had a bigger significance next to prior experience in all the models. Though the fuel cost is controlled, the initial investment costs of electric vehicles are higher. This is due to the unavailability of raw materials for the battery that in turn adds to the import expenses. To reduce that, the vendors and government should think of alternatives to Lithium-ion batteries which would be cost effective.

To promote electric vehicle sale, Government is providing schemes like FAME which provides subsidies and incentives to buy electric vehicles, but most people are not aware of those. Government can take steps in creating awareness regarding these policies and making it clear to the people. Also, making people understand the environmental consequences of the current vehicular population and the emissions are important to make them understand why sustainability is inevitable. The aesthetic values of the electric vehicles had a positive impact on the willingness. The electric vehicles can be redesigned to make it more appealing and rolled out in different colour variants to catch attention of customers. Current conventional vehicle usage was also another significant predictor, with people not owning a conventional vehicle more likely to prefer buying an electric vehicle. Maiden riders can be approached to try out the electric vehicles before they could start with conventional vehicles by giving a free demo. Through the sentiment analysis, it can be understood that people widely believe electric vehicles are environment friendly and reduce pollution. This is the unique selling point of the electric vehicles, and vendors can use it to promote the electric vehicle sales. People also described about the cost of the electric vehicles, and overall, they believe electric vehicles are 'good' as seen from the word cloud. Also from the word cloud, it is known that people are concerned about the charging stations, but still believe it is the future and is a good alternative to conventional fuels.

6 Scope and Limitation

The analysis part of the study covers the concepts of machine learning and sentimental analysis through text mining. Through the predicted results, the percentage of people who are willing to buy Electric Vehicles can be derived and the true intention of the people in purchasing Electric Vehicles in India can be known. The prediction models and the text analytics used in the study can be integrated to marketing analytics and people analytics. The model giving highest accuracy can in turn be adopted by Electric Vehicle industries by cascading several other Machine Learning or Artificial Intelligence models. Also, the research aims at predicting the customer willingness to buy generic Electric Vehicles and not their willingness to buy any specific type like Hybrid/Battery/Plug-in. Because of time constraints and COVID-19 pandemic, the sampling technique is chosen as convenience sampling. The probability or purposive sampling can be used to better represent the population. The constructs used for the study are totally based on the individual's perception towards environmental, subjective norm, aesthetic, cost and operational factors.

7 Conclusion

The different analyses made helped in understanding the significant predictors of the Electric vehicle purchase willingness, and they are the prior experience with the electric vehicles, cost convenience, aesthetics, gender, individual attitude and current conventional vehicle usage. Though government is initiating different schemes to raise funds for electric vehicles and promotes electric vehicles by giving subsidies, the government policy measure was not a significant predictor here. The study concludes that good aesthetics, male, individuals with sustainable thinking, maiden vehicle users and users who already owned an electric vehicle are the different attributes to be concentrated on to promote electric vehicles. The customers believe that electric vehicles can populate in future but are really concerned about the infrastructure and costs. If government and vendors bring in different solutions to combat these, the electric vehicles will come off with flying colours.

References

1. Afroz R, Rahman A, Masud MM, Akhtar R, Duasa JB (2015) How individual values and attitude influence consumers' purchase intention of electric vehicles—some insights from Kuala Lumpur, Malaysia. Environ Urban ASIA 6:193–211. https://doi.org/10.1177/0975425315589160
2. Ajex Thomas Varghese VSA and SVP (2021) A study on consumer perception and purchase intention of electric vehicles in India. Asian J Econ Financ Manag 4:13–25. https://doi.org/10.22214/ijraset.2021.35156

3. Bhalla P, Salamah I, Nazneen A (2018) A study of consumer perception and purchase intention of electric vehicles. Eur J Sci Res 149:362–368. Retrieved from https://www.europeanjournalofscientificresearch.com
4. Goutte C, Gaussier E (2005) A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. Lect Notes Comput Sci 345–359. doi: https://doi.org/10.1007/978-3-540-31865-1_25
5. Han L, Wang S, Zhao D, Li J (2017) The intention to adopt electric vehicles: driven by functional and non-functional values. Transp Res Part A Policy Pract 103:185–197. https://doi.org/10.1016/j.tra.2017.05.033
6. Jiang Q, Wei W, Guan X, Yang D (2021) What increases consumers' purchase intention of battery electric vehicles from Chinese electric vehicle start-ups? Taking NIO as an example. World Electr Veh J 12:1–19. https://doi.org/10.3390/wevj12020071
7. Kenan Degirmenci MHB (2017) Consumer purchase intentions for electric vehicles: is green more important than price and range? Resour Policy 7:1–10. doi: https://doi.org/10.1016/j.trd.2017.01.001
8. Kumar Shalender NS (2020) Using extended theory of planned behaviour (TPB) to predict adoption intention of electric vehicles in India. Environ Dev Sustain 23:665–681. https://doi.org/10.1007/s10668-020-00602-7
9. Lou Y, Wang W, Yang X (2017) Customers' attitude on new energy vehicles' policies and policy impact on customers' purchase intention. Energy Proc 105:2187–2193. https://doi.org/10.1016/j.egypro.2017.03.617
10. Nagaraj Navalgund GN (2020) Factors influencing purchase intention towards E-vehicles among the potential Indian consumers—a study on Karnataka region. J Soc Sci 23:551–563. https://doi.org/10.13140/RG.2.2.28103.11686
11. Okada T, Tamaki T, Managi S (2019) Effect of environmental awareness on purchase intention and satisfaction pertaining to electric vehicles in Japan. Transp Res Part D Transp Environ 67:503–513. https://doi.org/10.1016/j.trd.2019.01.012
12. Plötz P, Schneider U, Globisch J, Dütschke E (2014) Who will buy electric vehicles? Identifying early adopters in Germany. Transp Res Part A Policy Pract 67:96–109. https://doi.org/10.1016/j.tra.2014.06.006
13. RM A, Alenazi MJF (2021) A Comparative Study of Traffic Classification Techniques for Smart City Networks. Sensor 21:4677. doi: https://doi.org/10.3390/s21144677
14. Tanwir NS, Hamzah MI (2020) Predicting purchase intention of hybrid electric vehicles: evidence from an emerging economy. World Electr Veh J 11. doi: https://doi.org/10.3390/WEVJ11020035
15. Tu JC, Yang C (2019) Key factors influencing consumers' purchase of electric vehicles. Sustain 11. doi: https://doi.org/10.3390/su11143863
16. Xu Y, Zhang W, Bao H, Zhang S, Xiang Y (2019) A SEM-neural network approach to predict customers' intention to purchase battery electric vehicles in China's Zhejiang Province. Sustain 11. doi: https://doi.org/10.3390/su11113164
17. Ye F, Kang W, Li L, Wang Z (2021) Why do consumers choose to buy electric vehicles? A paired data analysis of purchase intention configurations. Transp Res Part A Policy Pract 147:14–27. https://doi.org/10.1016/j.tra.2021.02.014

A Data Science Approach in Quantitative Market Research

Mohammad Touhid Hossain Shakil and Rajkumar Rajasekaran

Abstract This research work aims to create a comprehensive marketing research model for the smartphone market that uses the data collected from the customers to support the decision process in framing the marketing program for any given company. At first, we have set the research objective and then constructed a questionnaire to collect the data through an online survey from the customers. As the research goal is preset and defined, we have adopted a descriptive research methodology for our research model. After identifying the research approach, we preprocessed the collected data and then applied the K-means clustering algorithm on the dataset to segment the market into different cluster numbers by buying capacity, loyalty, satisfaction level, and product ratings. We analyzed the market segments both quantitatively and qualitatively to draw valuable insights from them, which is then used to answer the marketing managerial questions to create and effective marketing program.

Keywords Market segmentation · Market research model · Smartphone market · Clustering · K-means · Decision support system

1 Introduction

The smartphone market is enormous [1] and has diverse customer groups that are distinct by demographic, geographic, psychographic, and behavioral attributes. Each of these groups has unique needs, characteristics, and buying behavior. These customers are homogenous inside the group but are heterogeneous to the other groups and hence they share a similar buying behavior [2]. As it is quite impossible to satisfy all customer segments of the market by offering the same value, the marketer first needs to identify and understand the needs of the different market segments along with several market metrics (market size, offerings by competitors, satisfaction and dissatisfaction of customers, potential opportunities, etc.) to create value

M. T. H. Shakil (✉) · R. Rajasekaran
School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India
e-mail: mthshakil02@gmail.com

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_31

for the customer segment [3] that would eventually capture the desired value in return from the customers. The advent of recent technologies and the generation of substantial digital data caused most of the companies including the smartphone market (a trillion-dollar worth market) to adopt a data-driven business strategy to get a competitive edge over one another. Market research is such a data-driven process of identifying customer needs, target customer segments, characteristics, behavior, and several other market properties which are then used to make decisions and to frame a customer-driven marketing program. This research work aims to design a comprehensive marketing research model for the smartphone market that can be used by marketers to identify customers, their needs, profitable market segments, competitors, internal and external factor affecting a particular market, and the marketing strategy that needs to be followed to become successful in the market.

1.1 Related Work

Although there are several research works related to the market segmentation, there exists no work that provides a comprehensive framework for the smartphone industry in particular that could be used by the smartphone marketers to create an effective marketing program. They have addressed problems that are too specific in nature and cannot be used in general to address marketing problems at large or for a broad range of problems. And among these works, for example, [4] used only demographic and behavioral attributes for segmentation, and [5] used psychographic, demographic, and behavioral attributes for segmentation. Reference [6] used only behavioral attributes for segmentation. References [7] and [8] used geographic and psychographic attributes for segmentation. Reference [9] used only psychographic attributes to segment the market. So, it is evident that the existing works only considered a subset of the segmentation variables to segment the market and are limited to narrowly addressing a specific part of a marketing problem. Only a few or none have addressed the smartphone market as a whole that can provide a unified market research model to guide the smartphone marketers to design an effective and holistic marketing program based on all the relevant attributes of smartphone market. In this marketing research model, we have devised a model that is generic in nature and encompasses a broad array of marketing problems in the smartphone market that includes all the segmentation types (demographic, psychographic, geographic, and behavioral) in its design. We collected data from more than hundreds of smartphone customers based on 26 critical attributes which are important in understanding the customers as a whole in framing an effective marketing program. We then used this data to segment the market and then analyzed them to draw useful insights by leveraging the power of Data Science tools and techniques to support the decision process for creating an effective marketing program.

2 Methodology

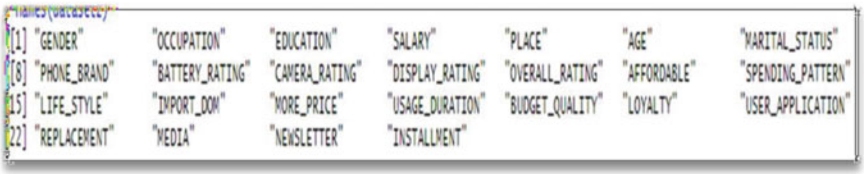
This marketing research model for the smartphone market comprises five fundamental parts: define research purposes, design research, collect data, prepare and analyze the data, and make recommendations for creating marketing programs. There are various methodologies to carry out market research, and this variation differs by the types of market research, research purposes, data sources, data collection methods, and data analysis techniques. Our methodology is descriptive because we have planned to explore and explain a particular marketing situation, i.e., the smartphone market [10]. After we have defined the research approach and specified the purpose, we then created a questionnaire to frame the research questions to conduct an online survey to collect data about smartphone customers. Few of the purposes include questions about the market's size, competitor's offerings, the market leader and loser, profitable market segments, buying capacities, advertising channels, product ratings, customer satisfaction and dissatisfaction points, etc. We have used them to draw insights from the market segments to make decisions about pricing, branding, product, advertising channels, and many others to create an effective marketing program. As the survey is done online, a digital form is created using Google Form because people are more familiar with Google Form and the survey Unified Resource Locator (URL) can be easily shared on online platforms like WhatsApp, Facebook, and other social and Internet media that have a huge number of smartphone users [11]. The survey is conducted in Bangladesh as this research intends to target the smartphone customers in Bangladesh to create the market research model. After conducting the survey, we collected the data and downloaded it in the Comma-separated values (CSV) form. Afterward, we have used R to preprocess and analyze the data. R programming language is one of the most preferred languages for conducting academic research and is best known for statistical data analysis having rich libraries for all sorts of statistical analysis [12]. As the research goal is to build a model rather than focusing on coding, we can leverage rich R libraries that are made for almost all heavy data analysis tasks. We used the *Tidy-verse* [13] library for data manipulation to fill the missing values, data standardization, scaling, splitting, and encoding of data. After preprocessing the data, the dataset is divided into several clusters in order to segment the smartphone market and draw valuable insights to understand the smartphone market and to create an effective marketing program for any given companies. As our research approach is descriptive in nature, we use *the K-means algorithm* [14] to segment the market [15].

In marketing segmentation, there are very few algorithms like K-means that could counter it when it comes to process a large dataset in shortest amount of time and simplicity. *The Factextra library* [16] is used to visualize the data. We have used Silhouette to obtain the optimal number of clusters for market segmentation. Silhouette is one of the most popular methods for determining the optimal clusters. This technique computes the average Silhouette for different values of k (cluster number). For a range of k values, the one that produces the highest average Silhouette is taken as an optimal cluster number for the dataset.

3 Results and Discussions

In the research-purpose definition phase, we have created twenty-six questions that cover almost all the important attributes of demographic, geographic, psychographic, behavioral, and marketing mix-related data of the customers in the smartphone market. We have put these questions into a Google Form and conducted the survey online. After completing the online survey, we gathered and then converted the data into a single CSV file. The CSV file comprises twenty-six columns, i.e., variables Fig. 1, each representing a research question. We loaded the CSV file containing the survey responses into the R environment. We cleaned the encoded dataset Fig. 2 by removing and filling up the missing data using Tidy-verse Library. In the encoding process Tables 1 and 2, we have also encoded the nominal data, but practically it makes no sense when it comes to clustering. So, nominal data is exempted from being considered as clustering variables.

We know from the real-world marketing problem in the smartphone industry, the most important considerations are price, affordability, loyalty, overall customer



```
[1]  "GENDER"       "OCCUPATION"     "EDUCATION"     "SALARY"         "PLACE"          "AGE"         "MARITAL_STATUS"
[8]  "PHONE_BRAND"  "BATTERY_RATING" "CAMERA_RATING" "DISPLAY_RATING" "OVERALL_RATING" "AFFORDABLE"  "SPENDING_PATTERN"
[15] "LIFE_STYLE"   "IMPORT_DOM"     "MORE_PRICE"    "USAGE_DURATION" "BUDGET_QUALITY" "LOYALTY"     "USER_APPLICATION"
[22] "REPLACEMENT"  "MEDIA"          "NEWSLETTER"    "INSTALLMENT"
```

Fig. 1 Variables list in the dataset

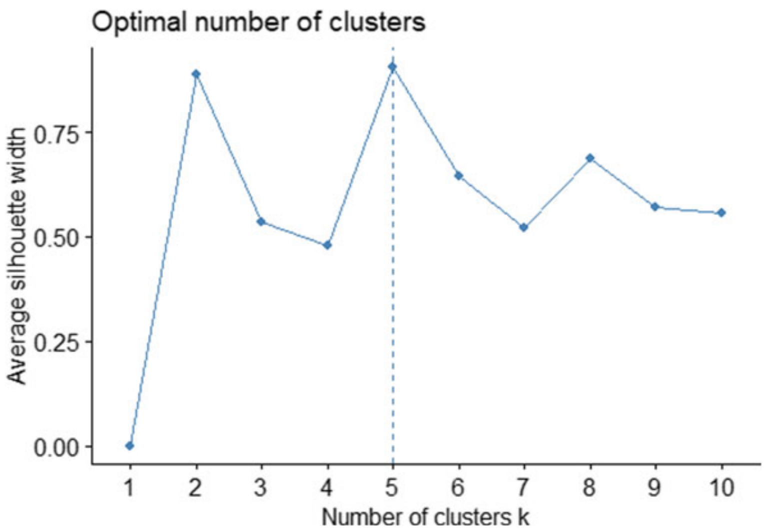


Fig. 2 Optimal cluster numbers using Silhouette method

Table 1 Numerical encoding of variables part 1

Gender	Occupation	Education	Place	Numerical encoding
Female	Government service	Domestic	Luxurious	1
Male	Others	Imported	Moderate	2
	Private jobs		Simple	3
	Students			4

Table 2 Numerical encoding of variables part 2

Marital status	Spending pattern	Import/domestic	Life style	Numerical encoding
Married	Spends lot	Domestic	Luxurious	1
Unmarried	Spends moderately	Imported	Moderate	2
	Saving person		Simple	3

satisfaction, and the features of the smartphones. If a customer likes the features of a smartphone and can afford it, they tend to buy it. If a customer has already used a particular smartphone and has a good experience with it, the chance of purchasing the same smartphone or smartphones of that brand again becomes higher. Therefore, we have selected the variables that are related to battery, display, loyalty, camera, and price to segment the market because these are the variables that predominantly shapes the user experience of a user while they uses a smartphone and affect the users in making the decisions to buy a smartphone. After selection of the segmentation variables, we created a separate dataset only containing data of the segmentation variables from the encoded dataset for market segmentation. Before moving to the clustering, the ideal number of clusters for the dataset is determined first. Figure 2 shows the ideal number of clusters is five according to the Silhouette method.

We segmented the market into five clusters. We have two figures that describes the clusters, one visually Fig. 3 and the other with textual data Fig. 4.

We have segmented the market with an optimal number of clusters for the dataset. We have visualized the segmentation and have data from the clustering to be analyzed. We have assigned cluster number to all the data rows in the dataset to which they belong. Then we separated all data rows clusterwise. We have total five clusters and hence five different datasets, each of these represents a segment of the market.

3.1 Analysis of Patterns Obtained from the Whole Market After Segmentation

Before analyzing a particular segment of the market, let us first interpret the patterns that are visible after the segmentation of the market. We see there is a positive

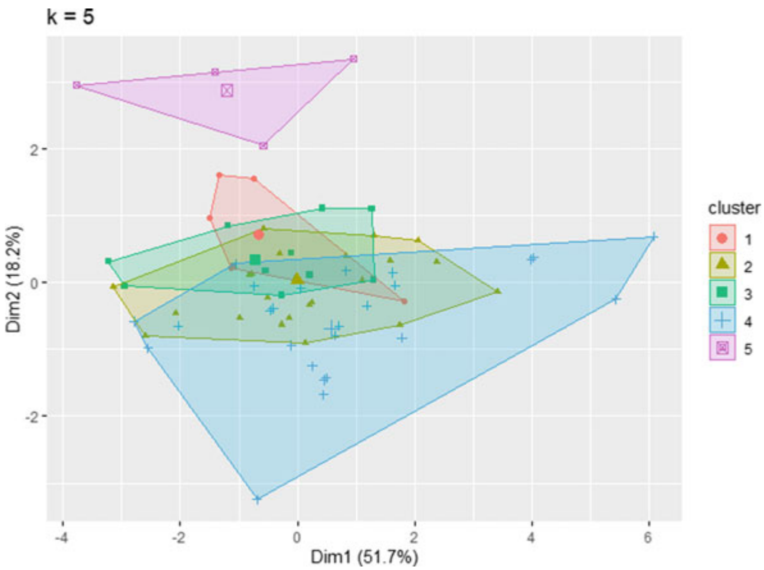


Fig. 3 Visualization of clustering (market segmentation) using K-means algorithm

```
> print(k5)
K-means clustering with 5 clusters of sizes 12, 6, 4, 30, 37

Cluster means:
  dataset2.CAMERA_RATING dataset2.BATTERY_RATING dataset2.OVERALL_RATING
1               3.750000                4.000000                4.000000
2               3.666667                4.333333                3.833333
3               4.000000                4.000000                4.000000
4               2.700000                3.766667                3.633333
5               3.648649                3.513514                3.810811
  dataset2.DISPLAY_RATING dataset2.LOYALTY dataset2.AFFORDABLE
1                4.083333         3.750000            30000.00
2                4.333333         3.500000            43333.33
3                4.250000         3.250000           100000.00
4                3.666667         2.866667            13000.00
5                3.864865         3.324324            20000.00

Clustering vector:
 [1] 5 4 4 1 2 5 2 5 4 5 4 2 5 1 5 3 5 3 5 4 1 5 5 5 4 1 4 5 4 5 5 1 3 1 2 5 5
[38] 5 5 4 4 5 4 1 5 2 1 3 2 4 4 5 1 4 4 5 4 5 5 4 1 5 5 5 4 4 5 4 4 1 5 5 5 4
[75] 4 4 1 4 4 4 5 5 4 4 5 5 5 4 5

Within cluster sum of squares by cluster:
[1] 4.141667e+01 1.333334e+08 1.750000e+01 1.800002e+08 1.277838e+02
 (between_SS / total_SS =  99.0 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

Fig. 4 Segmented smartphone market

correlation between price, product features, and brand loyalty. Relatively expensive smartphones have better product ratings and customers are more loyal to those phones. Most of the smartphone users' (89%) buying capacity is below 20 k Taka.[1] So, marketers can use this information to determine what type of smartphones in terms of price range they will offer in the market. From the data, it is quite evident, if they offer expensive phones there won't be enough customers to buy them so they can produce them only on small scale. Using the segmentation, marketers can identify themselves exactly where they belong and after that, they can further analyze the segment individually to draw insights from them.

3.2 Analysis of an Individual Market Segment

The individual cluster-wise dataset can be used to answer all marketing-related managerial questions, analyze a particular market segment, and make a comparison to make an effective marketing decision. We have selected cluster four, Fig. 5, because it contains the highest percentage of customers among all the market segments. Cluster 4's average camera rating is 2.7, loyalty is 2.8, battery rating is 3.7, display rating is 3.7, overall rating is 3.6, and the buying capacity is around 13 k Taka. We further segmented, Fig. 6, the users of this segment by different variables to get more detailed insights about it. Cluster 4's average camera rating is 2.7, loyalty is 2.8, battery rating is 3.7, display rating is 3.7, the overall rating is 3.6, and the buying capacity is around 13 k Taka. We further segmented this segment by different variables to get more detailed insights about segment four. We see that Samsung (30%) tops the segment, followed by Xiaomi (21.6%). So, customers in this segment prefer to buy these phones and have a good rating on the product. So, we can tell that Samsung is the market leader in this segment. Lava and Honor have a shallow presence (3%) and can be named as losers. In this segment, users' primary purpose for smartphones is general use (70%) follow by photography (22%). They are interested in buying a phone with easy installment (55%). The majority of customers in this segment live in urban areas (76%). Their lifestyle is mostly simple and moderate (89%), and more than half of them are using their current phones for more than a year. They watch social media, YouTube, and television the least. Most of the customers lead a simple and moderated lifestyle, this says they belong to the middle class. The majority (92%) of the customers in this segment are either studying in a college or have already graduated. 97% of the customers fall in the age group of (16–35) or 89% are men.

Now, if we interpret the analysis from a marketing point of view, it will help us to relate all these insights to frame a more effective marketing program for any new product or an existing product in the market. As the buying capacity of this market segment is around 13 k Taka and most of the customers belong to the middle

[1] 1 US Dollar equals 85 Taka.

```
> head(cluster5)
  X.1  X GENDER OCCUPATION EDUCATION SALARY PLACE AGE MARITAL_STATUS
1   1  1      1          4         1  10000     3  25              2
2   6  6      2          4         1  10000     3  25              2
3   8  8      2          4         1  10000     3  25              2
4  10 10      2          4         1  10000     3  25              2
5  13 13      2          4         1  10000     3  25              2
6  15 15      2          4         1  10000     3  25              2
  PHONE_BRAND BATTERY_RATING CAMERA_RATING DISPLAY_RATING OVERALL_RATING
1          14              3             4              4              4
2          14              3             3              3              3
3          12              5             5              5              5
4          14              4             1              3              3
5          14              3             4              4              4
6           9              5             4              4              4
  AFFORDABLE SPENDING_PATTERN LIFE_STYLE IMPORT_DOM MORE_PRICE USAGE_DURATION
1      20000                2          1          2          3              3
2      20000                2          1          1          2              4
3      20000                2          2          2          3              5
4      20000                3          3          2          1              1
5      20000                2          3          2          1              1
6      20000                3          3          1          2              1
  BUDGET_QUALITY LOYALTY USER_APPLICATION REPLACEMENT MEDIA NEWSLETTER
1              2       5                3           2     3          3
2              1       4                3           3     2          1
3              2       5                3           2     3          2
4              1       3                3           3     2          2
5              1       2                3           2     2          2
6              2       3                4           3     3          2
  INSTALLMENT cluster
1           3       5
2           3       5
3           1       5
4           3       5
5           2       5
6           2       5
```

Fig. 5 Cluster 4 (segment four of the market)

class, marketers can attract these customers with discounts and offer to offer the best low-budget smartphones.

They can also offer these low-budget phones through installments as most of the customers in this segment are interested in buying a phone through installments. Customers in this segment watch YouTube and social media the most, television and newspaper the least. For advertisement, marketers can target this customer group through social media and YouTube as these platforms have the tools to target customers by age, occupation, gender, location, etc.

In this way, marketers can save a lot of money avoid investing in the wrong advertisement channels for a particular customer group. This information can also be leveraged by salespeople to understand customers well, e.g., which customer to offer which products. As most of the customers in this segment live in urban areas (76%), marketers can prioritize their distribution channels in urban areas. We see, despite having a moderate battery, display, and overall rating, the loyalty rating is quite inconsistent. The strongest reason could be the low camera rating which is 2.7. In today's world of social media, customers prefer smartphones that have a very good camera to click high-resolution pictures and shoot videos. So, smartphone marketers of this segment can improve the battery performance to increase their overall rating

BRANDS:	Apple	Honor	Huawei	Motorola	One Plus	Real Me
USER(%):	11	3	11	3	8	5.5
BRANDS :	Samsung	Symphony	Vivo	Xiaomi		
USER (%):	30	3	5.4	21.6		
AGE RANGE:	< 25	25-35	> 45			
USER(%):	62	35	3			
EDUCATION LEVEL:	College	High School	University			
USER (%):	46	8	46			
PLACE:	Rural	Sub-Urban	Urban			
USER(%):	8	16	76			
MEDIA TYPE:	Social Media	YouTube				
USER(%):	49	51				
GENDER:	Female	Male				
USER (%):	11	89				
LIFE STYLE:	Luxurious	Moderate	Simple			
USER (%):	11	38	51			
USER APPLICATION:	Business	General	Photography			
USER(%):	8	70	22			
MORE PRICE:	MAYBE	NO	YES			
USER (%): 4	3	22	35			
USAGE DURATION (LESS THAN):	1YEAR	2YEAR	3YEAR	4YEAR	5YEAR	
USER(%):	46	32	11	11	3	
INSTALLMENT:	No	Maybe	Yes			
USER (%):	24	21	55			

Fig. 6 Further segmentation of segment four

and particularly loyalty to retain the existing customer base. A new marketer who wants to create a new product for this segment can also make their products' battery better to gain upper hand in the market. As Samsung is the market leader, other marketers can follow Samsung very closely to understand their recipe of success and can also follow the market loser like Honor to avoid making the same mistakes they do. Like the above marketing decisions, several other decisions and market insights can be drawn by segmenting the marketing with different combinations of variables available in the dataset.

4 Conclusion

The smartphone market is segmented based on the collected dataset from the users. The same dataset can be used to understand the market as a whole. Our main objective in this research work is to build a model using which smartphone companies can understand the market rigorously to make marketing decisions. Marketing research is very critical in this age of information because there are billions of users and thousands of competitors and it is hardly possible for any company to dominate the whole spectrum of the market. If we look at the data, we will find that some of the companies are best in one segment and performing below average in another segment. Even though a company can be equally good in all segments, it must create different marketing strategies for different market segments according to the market demands to hold its position [17]. In this research, we have built a marketing research model using the Data Science approach to analyze data quantitatively. For every step in the design of the research, we have applied the best and suitable principles so that this model helps the marketer to make marketing decisions specifically for the smartphone market. We segmented the market into five segments, which is the ideal number for our research problem. We analyzed the clusters, in general, to gain insights from them and also analyzed one segment individually to answer some of the managerial questions that are very important to create a marketing program for any given company. Although it is possible to draw n number of inferences and use the research model to answer a lot of managerial questions, we did not do that due to the scope of this work. We have seen that this model is not only limited to one aspect of the marketing program. It tried to capture all the variables that play a vital role in customers' buying decisions and need to be understood to create a useful marketing model. It has covered all the four Ps [18] of marketing (product, pricing, promotion, and place), which makes this research work stand out from other existing work. Moreover, most importantly, the dataset is collected from the real world (smartphone market in Bangladesh) without any research bias. This research shows how marketing and Data Science can be combined to build a model that entirely changes the decision support system for marketing. Earlier, market research had many limitations, but with the advent of new data analysis technologies, the entire process had become faster and more automated. Interdisciplinary research is being conducted like this research work to discover new avenues and how the latest technologies can be leveraged to solve real-world problems like marketing, which can be solved much more effectively, smoothly, and less expensively.

This research work can be improvised a lot using more sophisticated data collection methods and by collecting more customer data from the market. As researchers have restrictions and limitations in accessing and collecting the massive amount of data, marketers can collaborate with the researchers by giving access to their data and can provide funds so that more data can be collected. Besides, the marketers can come forward to create a digital infrastructure so that a massive amount of marketing data can be analyzed fast. Besides collecting the data only from the customers, other data sources can be used to build a model that supports and validate the model. For example, sales data can be used to validate the marketing decision made using this market research model. This market research model can be used as a decision support system for making marketing decisions in the smartphone market in Bangladesh. Likewise, data from the different markets can be collected then can be used for them as well. This research work leveraged the power of interdisciplinary research by combing marketing and Data science to create an effective market research model.

In future development, data collection processes extend beyond an online survey; more segmentation algorithms can be used to make the model robust and more accurate. Besides, a descriptive model, predicting models can also be built. We have the vision to make a complete marketing intelligence model for the smartphone market.

References

1. Intelligence M (2021) Smartphones market—growth, trends, COVID-19 impact, and forecasts (2021–2026). https://www.mordorintelligence.com/indus-try-reports/smartphones-market
2. Favero L, Belfiore P (2019) Data science for business and decision making. Academic Press, London
3. Kotler P, Armstrong G (2018) Principles of marketing. Pearson, London
4. Okazaki S (2006) What do we know about mobile Internet adopters? Cluster Anal Inf Manage 43(2):127–141
5. Mazzoni C et al (2007) Consumer behavior in the Italian mobile telecommunication market. Telecommun Policy 31(10–11):632–647
6. Chai Y et al (2017) Determining the effects of marketing mix on customers' purchase decision using the grey model GM(0,N)—case study of the western style coffeehouse chains in Vietnam. In: 2017 International conference on system science and engineering (ICSSE), IEEE, pp 575–578
7. Mao Z (2020) Research on marketing strategies of smart phone among college students. In: 2nd International conference on economic management and model engineering (ICEMME), IEEE, pp 178–182
8. Gide E, Zubaydi M (2019) Australian mobile user segmentation based on usage and purchasing measurements. In: 2019 IEEE Asia-Pacific conference on computer science and data engineering (CSDE), IEEE, pp 1–6
9. Estrin D et al (2010) Diversity in smartphone usage. In: MobiSys'10: proceedings of the 8th international conference on mobile systems, applications, and services. Association for Computing Machinery, New York, pp 179–194
10. Kumar V et al (2018) Marketing research. Wiley, New York
11. Global social networks ranked by number of users, https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/
12. Demidenko E (2021) Advanced statistics with applications in R. Wiley, New York
13. Wickham H (2019) Welcome to the Tidyverse. J Open-Sour Softw 4(43):1686
14. Piech C, https://stanford.edu/~cpiech/cs221/handouts/kmeans.html
15. Blattberg R et al (2008) Database marketing. Springer, New York
16. Mundt F, Kassambara A, Factoextra: extract and visualize the results of multivariate data analyse. https://cran.r-project.org/web/packages/factoextra/index.html
17. Kroesen M et al (2014) Mobile customer segmentation based on smartphone measurement. Telematics Inform 31(2):220–227
18. Lee Y et al (2018) Smartphone user segmentation based on app usage sequence with neural networks. Telematics Inform 35(2):329–339

Knowing Your Customers Using Customer Segmentation

Parichay Das and Vijendra Singh

Abstract This paper presents a customer segmentation review study to know more about the customer, and every successful business starts with knowing its customers. Trying to understand the customers on all levels is essential for companies. The main objective in customer segmentation is to understand their straight requirements, and business is to transform themselves into a crucial business segment to satisfy all their needs. It is vital to retain present customers and attract new ones in this viable market; understanding the customer is very important from this perspective. Different organizations are using the customer segmentation method, analyzing the value of the customer and providing them with better service according to customer need. This review studies the customer segmentation models and the most popular model recency, frequency, and monetary (RFM) to segment customers according to business needs, and segmenting customers according to profitability and seniority defines the customer journey. Also, highlighted are the different types of customer segmentation, customer segmentation models, and customer segmentation applications, which provide great insight into customer segmentation.

Keywords Customer segmentation · Customer segmentation model · Rrecency frequency and monetary (RFM) · Cluster analysis

1 Introduction

Using a customer segmentation model can increase profits for a business when there is increased competition among firms. It is more important to retain existing customers than to acquire new ones. Segmentation can help businesspeople customize marketing plans, identify trends, develop products, create advertising campaigns, and deliver relevant products to specific customers. Segmenting customers allows marketers to target their messages to their intended audiences better. Earlier purchase behavior, age, gender, location, income, and lifestyle are the most commonly used

P. Das (✉) · V. Singh
School of Computer Science, University of Petroleum and Energy Studies, Dehradun, India
e-mail: dasparic@gmail.com

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_32

attributes in customer segmentation [1]. Information gathering and dissemination on the Internet constitutes a critical component of eliciting interest from potential customers to engage in commercial relationships. With the ever-growing ability for e-commerce storage and processing, as well as the newest data mining and warehousing techniques, the online shopping process offers unprecedented opportunities for capturing knowledge [2]. A significant trading pattern among e-businesses is online shopping. Customer purchase behaviors change dynamically in such an online environment [3]. The researcher identified that customers are valuable to a company for which customers require promotion, and RFM values are unique. Competition among firms to remain in business has dramatically increased over the last few years. Retaining customers is crucial for business growth [1]. RFM is also helpful in determining the 20% of customers who makeup 80% of revenue [4]. Segmenting customers enables marketing strategies to be tailored specifically for each group based on their characteristics, allowing for more targeted marketing [5].

In this paper, the arrangement of the remaining sections is as follows: Sect. 2 describes a different type of customer segmentation. Section 3 presents and explains the customer segmentation model and model type. Section 4 offers state of the art in machine learning techniques for customer segmentation—Sect. 5 customer segmentation application and implementation. Section 6 concludes with some avenues for future research.

2 Type of Customer Segmentation

There are many business reasons to use these models, ranging from simple to complex. The following are some standard segments.

2.1 Demographic

For creating and delivering content based on that customer segment, many companies identify gender at the very least. When it comes to demographic and socioeconomic data, the collection and analysis are relatively straightforward. Its not difficult to gather the information and measure regarding demographics and socioeconomics. Statistical data on demographics and socioeconomic factors is relatively straightforward to collect and measure. Descriptive methods based on a priori variables are the best. There are several variables in this list, including age, gender, household size, household income, occupation, education, religion, and nationality. A segment, retailers, is most likely to combine these variables with others. Socioeconomic status, educational level, and household income also show good correlations [6].

2.2 Behavioral

As a result of past observed behavior, future actions can be taken thoughtfully, such as purchasing from a particular brand, purchasing for certain occasions, or any significant events such as birthdays, anniversaries, getting married, having a baby, and moving somewhere. It is essential to consider the reasons for purchasing customers' products, and if those reasons might change throughout the year, it is an indication that their needs are changing. Essentially, consumer behavior refers to how people react to a company's marketing initiatives. In addition to cultural and psychological factors, where, when, and why consumers purchase is also influenced by these factors. Some practitioners recommend behavioral segmentation to identify the characteristics of a particular segment [7].

2.3 Psychographic

Psychographic segmentation includes softer measures like attitudes, beliefs, or even personality traits [8]. It is a strategy that predicts what market segments a product will appeal to based on respondents' lifestyles, personalities, hobbies, and socioeconomic classes—Segmentation of respondents based on their attitudes, interests, and opinions [7].

2.4 Social Media Segmentation

Among the many uses of social media, it has excellent potential as a segmentation method. By segmenting consumers based on their behavior on social media, marketers can conduct instant and targeted marketing [7].

3 Customer Segmentation Models

This section presents different customer segmentation models, including recency, frequency, monetary (RFM model), high-value Customer (HVC) model, and customer status model.

3.1 Recency, Frequency, Monetary (RFM)

Identification of high-value Customers (HVC) is another method of RFM method. RFM is frequently used in the retail and banking space to identify customers based on their last purchase (recency), how often they make purchases (frequency), and how much money they spend (monetary). Recency, frequency, monetary (RFM) analysis: Hughes of the American Database Institute proposed the RFM model in 1994. The RFM tool is also commonly used for determining customer lifetime value, customer segmentation, and behavior analysis. For customer value analysis, RFM model is also used, and researchers have extended it in different ways. Studies of customer behavior reveal that it is more likely that purchasers will buy from the seller again if the R or F is high.

Additionally, the more M a customer has, the more likely it is that customer will buy from the seller again [9]. In database marketing, recency, frequency, and monetary are very powerful and recognized techniques. Based on the customers prior purchasing history, rank the customer's RFM is also used. Large applications involving many customers like online purchases and retail products recommendation primary method is RFM. Recency, frequency, and monetary are the three dimensions of the customer of this method.

- **Recency (Based on the Customer Purchase Count)**: The recency value refers to the number of days between a customer's two purchases. A smaller value of recency means customers visit the company repeatedly within a brief period. Similarly, a significant value of recency means the customer does not see the company frequently.
- **Frequency (Based on the Counts of Purchases by Customers)**: Frequency is measured by the number of purchases a customer makes over a specified time period. The higher value of the frequency means the customer is more loyal to the company.
- **Monetary (Based on the Money Spend by the Customer)**: Within a certain period, the money spent by the customer for purchasing the products is called Monetary. When the customer pays a higher amount of money for the company, it increases its revenue [1] (Table 1).

3.2 High-Value Customer (HVC)

Small, medium-sized, and large businesses all know where high-value customers come from. They also know which characteristics these customers share so that they can attract more of them.

Table 1 Comparative analysis of recency, frequency, monetary models from the literature

Authors	Location	Business area	Methodology	Key finding
Sarvari et al. [10]	Turkey	Restaurant chain	Association rules + RFM analysis, K-means clustering	RFM variables affected by rule associations must be weighted appropriately with an appropriate segmentation approach and segmentation approach
Christy et al. 1]	The U.S	Online retail	K-means + RFM analysis + fuzzy C-means	The researcher can compare the storage, synchronization, and execution times of clusters based on different customer characteristics
Liu et al. [11]	Korea	Retail	Association rules + RFM analysis + K-means	CRM strategies based on customer segments and data mining techniques to identify purchasing patterns using VIPs' transaction data
Dogan et al. [3]	Turkey	Retail	K-means + RFM analysis	The purpose of customer segmentation without considering the customer's expenses is to understand the customer better, create well-designed strategies, and make more efficient decisions

3.3 *Customer Status*

Based on the customer's last purchased item or engaged order, most companies categorize them into active and lapsed customers. Active customers have purchased any non-luxury products in the past 12 months, whereas lapsed customers have not made any purchases in the past 12 months.

4 Machine Learning Techniques for Customer Segmentation

Unsupervised learning is an essential tool for customer segmentation. To target the right audience, companies can use clustering techniques to identify different segments of customers [32].

4.1 K-Means

Many of the researchers also use the K-means algorithm for customer segmentation using cluster analysis. In this algorithm, every cluster is connected to a central point (mean of points) using a sorting method called K-Means. A cluster, usually Euclidean distance, assigns every object to a cluster from its closest centroid. At the beginning of the simulation, defining K clusters and selecting their initial centroids at random should be done [1].

Customer recency, frequency, money (RFM) values have categorized the segmentation of enterprise customers with similar behavior into different segments. Analyzing a company's transactions over a defined period, segmentation helps the companies identify the potential customers and understand the needs of the customers. The importance of retaining existing customers outweighs finding new ones. Marketers are deploying marketing strategies tailored to the needs of a specific segment to maintain customer retention. RFM is first performed on the transaction data to extend the K-means algorithm to the cluster. K-means is considered to be a novel approach since it chooses the initial centroids. The obtained results of the different methodologies are compared by their iterations, cluster compactness, and execution time [1].

Marketing campaigns that aim to improve businesses and increase revenues should consider customer segmentation. A clustering algorithm can assist marketing experts in achieving their goals. Advanced analytics techniques are becoming increasingly important with the spread of data warehouses and high-dimension databases such as customer relationship manager (CRM). The investigation focuses on comparing data analytics algorithms using TIC CRM data, specifically the K-means and SOM. Promising clustering results are shown by K-means, in the sense of speed, quality of clustering, and visualization, SOM has outperformed. Analyzing customers' interests through segmentation is useful. It is primarily intended to provide proof of the principle of big data analytics (based on a small data set): how big data analytics is used [13].

4.2 Support Vector Machine

Segmentation of customers based on support vector clustering, a technique that has remained in the academic spotlight for a long time, has been proven to achieve good results. The technique is unlike the more traditional hierarchical cluster analysis. Accordingly, the findings indicate that this methodology can reduce customer data heterogeneity, providing managers with improved cluster segmentation analysis. The support vector machine (SVM) is a set of algorithms that can determine consumer segments using support vector clustering (SVC) based on available data. Consequently, a theoretical framework for customer segmentation and machine learning [14].

4.3 Regression

A predictive model is constructed by identifying insured/entities characteristics and modeling the threat of that insured/entity to the insurer. Regression analyses often use generalized linear models (GLMs) to predict risk. Its simplicity in interpretation makes linear regression an essential tool in the insurance industry for modeling risk. However, it is not uncommon to focus on binary outcomes, i.e., whether an event has occurred or not, for example, the likelihood of a fraudulent claim, the propensity to attrition, etc. Typically, when considering cases like this, logistic regression is employed to estimate the probability of occurrence of a given event, such as predicting the likelihood of a fraudulent claim or predicting customer retention/persistency [15].

4.4 Random Forest

Random forests are also helpful when clustering, which is extremely popular for regression and classification. By using a proximity matrix as input directly, the spectral clustering method can utilize information. Converting proximity to distance, multidimensional scaling is possible by transforming observations x observations into observations x dimensions. Data in this format can be clustered as usual using k-means and model-based clustering. When data is mixed (as it often is in marketing settings), random forests provide an efficient, flexible method for clustering analysis. A bonus is that you can enhance the relevance of the segmentation output by supervising the clustering with a target vector of interest (such as sales, category purchases, or income) [16].

4.5 Neural Networks

Organizing the market by segment plays an essential role in literature and software dealing with customer relationship management since customer satisfaction directly affects it. By labeling a group of customers with a particular label, it is common to categorize customers into two categories. Multilayer perceptrons (MLP) can be used to train and test machine learning models using data derived from potentially segmented customer bases within our company, along with statistically derived data. Training neural networks with similar features involves using extensive grid search algorithms to pick hyperparameters. The system achieved good generalization of our customer segmentation strategy and, therefore, a reasonable accuracy within a few epochs. The company's data framework systematically analyzes customer information and decides whether promoted or those who should remain unchanged. Automated decision systems can increase customer satisfaction [17].

With the advancement of information technology (such as scanners, cookies, and other electronic-based methods of collecting data), researchers are collecting unprecedented amounts of data about individual customers. As larger and more complex customer databases become, the methods for analyzing them become burdened and suffer. Segmenting retail databases using artificial neural networks (ANNs) is an alternative method. An evaluation and comparison of Hopfield-Kagmar (HK) clustering, based on Hopfield networks, is presented alongside K-means and mixture model clustering algorithms. Mixture models and K-means clustering algorithms could not be as useful for retailers as ANNs, which provide more homogenous segmentation solutions and are less affected by initial conditions[18].

It is challenging to visualize a dataset with many dimensions. The self-organized map (SOM) or Kohonen Map enabled us to simplify the presentation and explore meaningful relationships. Data is converted into two-dimensional spatial data by neural networks using SOM, which represent relationships between the data objects. The nodes on a map are ranked according to how similar they are to one another. The advantage of reducing the high dimensions to a map would be more accessible and visually appealing to visualize. Similarly, grouping data together would facilitate clustering [13]. According to Qadadeh et al. [13], SOM with K-means is more effective in customer segmentation.

Using traditional K-means and fuzzy C-mean algorithms, they performs a transaction-level RFM analysis on the transactional data. According to RFM analysis, further research has used various algorithms to cluster the customers. In the initial preprocessing, outliers are removed as well, as meaningful instances are filtered. Using the Z-core, data can be compared with its mean and standard deviation to identify outliers. It is mapped to 0 and 1 that mode and standard deviation are related. Then, three clustering algorithms are employed: K-means, fuzzy C-means, and repetitive median based K-means (RM K-means) [1]. According to Christy et al. [1], RM K-means are more effective in customer segmentation.

4.6 Optimization Methods

Cluster analysis is significant since there is no suitable number of clusters in real-world problems, so determining the number of clusters is imperative. Clustering based on automatic selection of clusters and automatically divides instances into those clusters is known as automatic clustering. The IABC algorithm combines improved artificial bee colony optimization algorithms with K-means to create an automatic clustering algorithm [19].

It essentially involves grouping items of similar nature together. Using multivariate analysis as a method is standard. A hierarchical solution, such as Ward's minimum variance, and nonhierarchical strategies, such as K-means, are used. As computer power and costs have increased, artificial neural networks (ANNs), distributed and parallel systems that solve information processing problems, have recently been applied to marketing. ANN and multivariate analysis are also discussed in this study, and the proposed approach is a two-stage approach. An initial self-organizing map is used to determine the number of clusters and the starting point, and then the K-means algorithm is applied to determine the final solution. A sophisticated analysis of consumer behavior can guide marketing analysts in developing the most effective marketing strategies. Case studies further help illustrate the applicability of the proposed method [20].

Many mature industrial markets have few applications for need-based or micro-segmentation, and this empirical study proposes a model to address that need. This integrated framework combines the needs-based (need-based) segmentation concepts of industrial buyers with consumer-based models. This study examined the supply chain for a complex commodity, using a segmenting model based on the motivations and benefits of industrial buyers. There was also an examination of descriptive data. Accordingly, this study recommended custom-designed product services' packages and communication strategies. Other industrial settings would be the subject of future research [21].

In an innovative approach to segmenting cross-national markets, this study proposes a TPA that combines statistical and data mining techniques. The first step of the study was to determine whether there was a difference between the various national clustering factors using a statistical method called multi-group confirmatory factor analysis (MCFA). A data mining analysis (a two-level SOM) is used to define the actual clusters within each nation in the second phase. Using a two-level SOM effectively reduces the complexity of the reconstruction task and noise. Study participants were Korean and Japanese players, representing the global gaming industry in terms of online gaming users [22].

The difficulty in analyzing bank databases for customer behavior management has to do with the fact that they contain monthly accounts and daily transactions. The model was based on data mining and behavioral scoring for managing credit card customers in a bank. Using recurrence, frequency, and monetary behavior scores to predict repayment behavior, self-organizing map neural networks are utilized to identify client groups. Furthermore, the same method was used to divide consumers into three major groups based on their profitability. Profiling then occurs in accordance with the customer's features determined by the associated rule. Study findings indicate that objectively identifying customers with a behavioral scoring model facilitates the development of marketing strategies [23].

Different researchers applied different AI algorithms for their customer segmentation problem likes K-means, Mini-Batch K-means, hierarchical clustering, gaussian mixture model, and DBSCAN most popular algorithm. Below, describing comparison study and executed using a customer dataset, a company that sells software wants to examine the effectiveness of its outreach efforts in attracting new customers or increasing consumption among existing customers.

Segmentation of the customers depends on the relationships with the customers. Enterprises place a higher priority on retaining existing customers than finding new ones. To begin with, they used RFM analysis and customer segmentation before employing K-means clustering, fuzzy c-means, and after modifying K-Means clustering, they developed the RM K-means algorithm. There is a clear difference in the execution time of the K-means algorithms, as well as the number of iterations for each algorithm. Segment customers according to their recent purchases, frequency, and financial values. It is determined that centroids are more reliable and are based on established medians of data distribution. Based on the customer's buying behavior, marketers can tailor their marketing strategy [1]. Companies need to understand their customers' data better in all aspects. Customers and companies are more engaged when they can detect similarities and differences, predict their behaviors, and offer better options and opportunities. Customer segmentation based on their data became essential in this situation. Companies have used the recency, frequency, and monetary (RFM) values for many years to identify which customers are valuable and which require promotional activity. Many organizations and individuals analyze their stored data using data mining tools and techniques. The data mining process facilitates the grouping of people, objects, and other data. The author proposes two different clustering models based on the RFM values of 700,032 customers in this paper [3].

Expected growth of consumer-oriented e-commerce is crucial to segment the online shopping market, analyze Internet customers' needs, and segment the market based on their expectations. It intends to do an exploratory segmentation of this market to shed some light on its structure and to analyze the growth of online shopping. The study applied the tandem approach to cluster-based market segmentation to real-world survey data on internet users. A factor analysis of the observable data and cluster analysis of the factor scores proves that this model is valid. The factors themselves, selected as the best and most reliable predictors of user propensity to buy online, support the validity of this model. Segments, which are defined based on profit, are therefore optimized for actionability [2]. Customer list segmentation with direct marketing, a centralized response model, such as the recency, frequency, monetary (RFM) model, logistic regression, and neural networks, can be used to target and segment your audience.

Nevertheless, if the database contains considerable customer heterogeneity, the models may be misleading. To take advantage of this heterogeneity, researchers have developed methods that combine two or more ways. Recently, literature has begun to respond to this challenge through direct marketing-related response models.

Many response modeling tools have been developed over the years to help direct marketers solve their targeting and prediction problems. Popular ones include (recency, frequency, monetary (RFM), logistic regression, and neural networks (NNs) [24]. The establishment of computers in business applications has led to computers storing a wide range of data. Particularly after the advent of database systems, data has been getting much more significant. The goal of any manager is to glean knowledge from this massive set of data. A customer-centric sales strategy nowadays will lead to higher sales because production tends to shift to customer-centric models. To explain a data set, the researchers used different assumptions and developed a model of expertise based on easily understood rules. Expert systems can also incorporate rules as a knowledge base. Answering some questions allows any user to take its advice. Some companies have already used such knowledge to provide managers with decision-making guidance [25].

Video information processing is one of the most challenging aspects of the database community since powerful processors, and large amounts of storage are needed. Presented here is an analysis of how clustering large video data sets, such as video clips, can be used as foundational work for subsequent processing, such as video retrieval. Multidimensional data sequences represent arbitrary sequences of video frames. Partitioning the data sequence into video segments that consider the temporal relationship of structures when laying out a plan, then similar clip segments are grouped. This paper presents a video segmentation and clustering algorithm that guarantees cluster quality satisfying predefined conditions via experiments performed with different video datasets [26].

Getting information on customer value is crucial for effective customer relationship management. Researchers have conducted numerous types of research to determine the customer lifetime value (LTV). The program has some limitations. Customers defecting is challenging to comprehend. In traditional cash flow forecasts, the previous profits of customers are taken into account [27].

An adequately segmented market is essential for an effective marketing and customer relationship management program. In the past, marketers used variables such as demographics and lifestyle to segment markets. Segmentation results are unreliable due to several problems, however, based on customer transaction histories, the authors propose a newly developed market segmentation methodology that utilizes variable-specific information, such as the products purchased and the monetary expenditures required to purchase these products. A purchase-based similarity measure, a clustering algorithm, and a clustering quality function are introduced in this paper. The closest customer purchase patterns are matched with each cluster using a genetic algorithm. The RFM model is then used to analyze the relative profitability of each cluster of customers following the segmentation process and also be discussed the findings of a practical marketing implementation study [28] (Table 2).

Table 2 Comparative analysis of customer segmentation techniques

S. No.	Authors	Customer segmentation models	Proposed methodology	Dataset and domain
1	Miguéis et al. [29]	Behavioral	VARCLUS algorithm	Retail lifestyle
2	Hamka et al. [30]	Psychographic	Context elicitation methods	Telecommunication
3	Tabi et al. [31]	Demographic	Sawtooth software analysis + choice-based conjoint analysis	Renewable energy
4	Brito et al. [32]	Behavioral	K-medoids and CN2-SD	Online fashion
5	Christy et al. [1]	RFM	K-means and fuzzyc-means algorithm	Retail
6	Qadadeh et al. [13]	Behavioral	K-means and SOM	Insurance
7	Ballestar et al. [33]	Behavioral	Analysis on SPSS	Online cashback
8	Calvo-Porral et al. [34]	Psychological	S-O-R paradigm	Food retail
9	Dogan et al. [3]	RFM	K-means algorithm	Online retail
10	Chen et al. [35]	Psychological	(ECLAT) algorithm	Hi-Tech
11	Dadzie et al. [36]	Demographic	Data mining	Country economies
12	Sheikh et al. [37]	RFM	K-means clustering algorithm and LRFMP model	FinTech
13	Wu et al. [38]	Behavioral	K-means algorithm	Online retail
14	Ferrera [39]	Behavioral	Data mining	Hi-Tech

5 Customer Segmentation Application

Utilizing machine learning algorithms to discover new customer segments is another way to segment customers [39].

5.1 Banking and Financial Services Fraud Detection

Machine learning methods have been successful in several areas in the classification, predictions of credit defaults and fraud, and forecasts of foreign exchange markets. Financial institutions can address their products and services to homogeneous customers through customer segmentation and use customer segmentation for the foundation for good business development [40].

5.2 E-commerce Cashback Model

Customers can be categorized based on their commercial activity or their position on a social networking site. Customers benefit from the affiliate websites they visit while connected through this social network. Cashback business models rely heavily on word-of-mouth recommendations to be successful as these recommendations enhance customer acquisition and strengthen loyalty among existing customers[33].

5.3 Social Media Marketing Planning

According to Julian (2018), customer segmentation is essential in industrial buying in the age of digital, social, and mobile marketing (DSMM). Another vital customer segmentation application is digital and social media marketing [41].

6 Conclusion and Future Scope

An analysis of customer segmentation detail is presented here. By considering the online purchase history of a retailer, this paper analyzes customer purchase habits systematically. But, most of the customer segmentation domain area of research is retail. Organizations segment their customers for their growth and also try to gain more customers from the market and entice them to become users, subscribers, and purchase products and services from the organization. Almost all the organization's revenue is generated through memberships, products, and services. Unfortunately, the data they are getting is unstructured and raw. Because of that, every segmentation of customers is influenced by the type of organization and its goals. This review studies covering customer segmentation, method of customer segmentation, techniques of customer segmentation, and customer segmentation application. A company is thinking about customer segmentation when it is concentrating on its growth. In the customer segmentation model, RFM model is mainly used. But retention of existing customers and increasing the number of new customers is quite tricky. It is the future scope of work, new methodology will be proposed for customer segmentation and will compare with the existing methodology.

References

1. Christy AJ, Umamakeswari A, Priyatharsini L, Neyaa A (2018) RFM ranking–an effective approach to customer segmentation. J King Saud Univ Comp Inf Sci
2. Vellido A, Lisboa PJG, Meehan K (1999) Segmentation of the online shopping market using neural networks. Expert Syst Appl 17(4):303–314

3. Dogan O, Ayçin E, Bulut Z (2018) Customer segmentation by using RFM model and clustering methods: a case study in retail industry. Int J Contemp Econ Adm Sci 8(1):1–19
4. Srivastava R (2016) Identification of customer clusters using rfm model: a case of diverse purchaser classification. Int J Bus Anal Intell 4(2):45–50
5. Ernawati E, Baharin SSK, Kasmin F (2021) A review of data mining methods in RFM-based customer segmentation. J Phys Conf Ser 1869(1):012085). IOP Publishing
6. Johnson TE, Danner L, Bastian SE (2017) Consumer segmentation methods for the wine market. Ref Modul Food Sci
7. Fu X, Chen X, Shi YT, Bose I, Cai S (2017) User segmentation for retention management in online social games. Decis Support Syst 101:51–68
8. LeBlanc D (n.d.) Creating actionable customer segmentation models. Google Cloud Blog. Retrieved 14 Aug 2021, from https://cloud.google.com/blog/products/data-analytics/creating-actionable-customer-segmentation-models
9. Mesforoush A, Tarokh MJ (2013) Customer profitability segmentation for SMEs case study: network equipment company. Int J Res Ind Eng 2(1):30–44
10. Sarvari PA, Ustundag A, Takci H (2016) Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis. Kybernetes
11. Liu RQ, Lee YC, Mu HL (2018) Customer classification and market basket analysis using K-means clustering and association rules: evidence from distribution big data of Korean Retailing Company. Knowl Manage Res 19(4):59–76
12. Team D (2019) Data science project—customer segmentation using machine learning in R. DataFlair. Retrieved 20 Oct 2021, from https://data-flair.training/blogs/r-data-science-project-customer-segmentation/
13. Qadadeh W, Abdallah S (2018) Customers segmentation in the insurance company (TIC) dataset. Proc Comp Sci 144:277–290
14. Albuquerque P, Alfinito S, Torres CV (2015) Support vector clustering for customer segmentation on mobile tv service. Commun Stat Simul Comput 44(6):1453–1464
15. Deloitte Consulting LLP, Wu P, Banerjee D, Nekkalapu KR (2014) Comparison of linear and logistic regression for segmentation. Deloitte Consulting LLP. https://www.casact.org/sites/default/files/presentation/rpm_2014_handouts_paper_2693_handout_1986_0.pdf
16. G (2018) Using random forests for segmentation—gradient. Medium. Retrieved 20 Oct 2021, from https://medium.com/gradient/using-random-forests-for-segmentation-e4793482f129
17. Ozan Ş, Iheme LO (2019) Artificial neural networks in customer segmentation. In: 2019 27th signal processing and communications applications conference (SIU). IEEE, pp 1–4
18. Boone DS, Roehm M (2002) Retail Segmentation using artificial neural networks. Int J Res Mark 19(3):287–301
19. Kuo RJ, Zulvia FE (2018) Automatic clustering using an improved artificial bee colony optimization for customer segmentation. Knowl Inf Syst 57(2):331–357
20. Kuo RJ, Ho LM, Hu CM (2002) Integration of self-organizing feature map and K-means algorithm for market segmentation. Comput Oper Res 29(11):1475–1493
21. Albert TC (2003) Need-based segmentation and customized communication strategies in a complex-commodity industry: a supply chain study. Ind Mark Manage 32(4):281–290
22. Lee SC, Suh YH, Kim JK, Lee KJ (2004) A cross-national market segmentation of online game industry using SOM. Expert Syst Appl 27(4):559–570
23. Hsieh NC (2004) An integrated data mining and behavioral scoring model for analyzing bank customers. Expert Syst Appl 27(4):623–633
24. Suh EH, Noh KC, Suh CK (1999) Customer list segmentation using the combined response model. Expert Syst Appl 17(2):89–97
25. Chen YT, Jeng B (2000) A customer segmentation mining system on the Web platform. In: AMCIS 2000 proceedings, Vol 35
26. Lee SL, Chung CW (2002) Hyper-rectangle based segmentation and clustering of large video data sets. Inf Sci 141(1–2):139–168
27. Hwang H, Jung T, Suh E (2004) An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. Expert Syst Appl 26(2):181–188

28. Tsai CY, Chiu CC (2004) A purchase-based market segmentation methodology. Expert Syst Appl 27(2):265–276
29. Miguéis VL, Camanho AS, e Cunha JF (2012) Customer data mining for lifestyle segmentation. Expert Syst Appl 39(10):9359–9366
30. Hamka F, Bouwman H, De Reuver M, Kroesen M (2014) Mobile customer segmentation based on smartphone measurement. Telematics Inform 31(2):220–227
31. Brito PQ, Soares C, Almeida S, Monte A, Byvoet M (2015) Customer segmentation in a large database of an online customized fashion business. Robot Comp Integr Manuf 36:93–100
32. Customer Segmentation: Definition, Models (2020) Optimove. Retrieved 19 Oct 2021 from https://www.optimove.com/resources/learning-center/customer-segmentation
33. Ballestar MT, Grau-Carles P, Sainz J (2018) Customer segmentation in e-commerce: applications to the cashback business model. J Bus Res 88:407–414
34. Calvo-Porral C, Lévy-Mangin JP (2018) From "foodies" to "cherry-pickers": a clustered-based segmentation of specialty food retail customers. J Retail Consum Serv 43:278–284
35. Chen H, Zhang L, Chu X, Yan B (2019) Smartphone customer segmentation based on the usage pattern. Adv Eng Inform 42:101000
36. Bock T, Uncles M (2002) A taxonomy of differences between consumers for market segmentation. Int J Res Mark 19(3):215–224
37. Sheikh A, Ghanbarpour T, Gholamiangonabadi D (2019) A preliminary study of fintech industry: a two-stage clustering analysis for customer segmentation in the B2B setting. J Bus Bus Mark 26(2):197–207
38. Wu J, Shi L, Lin WP, Tsai SB, Li Y, Yang L, Xu G (2020) An empirical study on customer segmentation by purchase behaviors using a RFM model and K-means algorithm. Math Prob Eng
39. Ferrera R, Pittman JM, Zapryanov M, Schaer O, Adams S (2020) Retailer's Dilemma: personalized product marketing to maximize revenue. In: 2020 Systems and information engineering design symposium (SIEDS), IEEE, pp 1–6
40. Smeureanu I, Ruxanda G, Badea LM (2013) Customer segmentation in private banking sector using machine learning techniques. J Bus Econ Manag 14(5):923–939
41. Müller JM, Pommeranz B, Weisser J, Voigt KI (2018) Digital, social media, and mobile marketing in industrial buying: still in need of customer segmentation? Empirical evidence from Poland and Germany. Ind Mark Manage 73:70–83

A Alzheimer's Disease Detection and Classification Using Customised Convolutional Neural Network

D. T. Mane, Divya Patil, Madhavi Sawant, Karina Maiyani, and Mehul Patel

Abstract Alzheimer's disease (AD) is regarded as acute diseases that cause death in humans, especially in those over the age of sixty, so it must be detected early in order to take preventative steps. In this paper, a supervised customised convolutional neural network (CCNN) is proposed for AD detection and classification using magnetic resonance imaging (MRI). AD is a form of dementia that causes memory, thinking and behavioural issues. Symptoms usually appear gradually and progressively worse. In the proposed system, we have used customised CNN classifier with five convolutional layers and five max pooling layers for the better classification of AD in four categories mild, moderate, very mild and none. Performance of the proposed CNN is evaluated on ADNI standard and results have obtained for multi-class classification of the disease. The proposed system gives the appropriate outcome for the classes mild, moderate, none, and very mild. The accuracy provided by proposed algorithm is 97.2% which is comparatively higher than the other existing algorithms.

Keywords Alzheimer disease detection · Convolutional neural network · Pattern recognition · Deep learning

1 Introduction

AD has become a widespread neurodegenerative brain illness in the elderly in recent years. As per study released by World Health Report, there are currently 44 million dementia patients in world, with that figure expected to rise to 76 million by 2030 and 135 million by 2050. AD accounts for 50–75% of these patients [1], and is characterised by a gradual onset and progressive deterioration of episodic memory [2]. Mild cognitive impairment (MCI) is a disorder that a person's reasoning skills deteriorate gradually but noticeably. MCI patients are more prone to becoming AD than those who do not have it. While there are no therapies that can cure AD, several medications have been utilised to slow down the course of specific symptoms and lessen

D. T. Mane (✉) · D. Patil · M. Sawant · K. Maiyani · M. Patel
JSPM's Rajarshi Shahu College of Engineering, Pune, Maharashtra 411033, India
e-mail: dtmane@gmail.com

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_33

the psychological impact on patients, such as memory loss [3]. As a result, correct diagnosis of AD patients or MCI at an early stage is critical in image processing applications, especially in the field of medical imaging. X-rays, CT scans and MRIs, to name a few, are imaging modalities that image body parts and provide a general view of internal organs. MRI pre-processing, segmentation, feature extraction and classification are the four stages of the Alzheimer's identification and classification processes, respectively. The main goal in the first stage is to remove any MRI noise that could cause problems. The extraction of the region of interest is done in the second level (Alzheimer region). The features associated with MRI images will be collected and processed in an image vector in the third stage, in preparation for the classification process, and finally the fourth stage, where classification will take place to specify the Alzheimer type.

This paper depicts the proposed system that uses CNN classifier to predict the stages of the patients suffering through AD. The paper contents are as follows. In Sect. 2 discussed related work that explains previous work of various researchers of AD detection and classification. Proposed system architecture contains working of CCNN algorithm along with mathematical model is presented in Sect. 3. The dataset details, loss and accuracy graphs, confusion matrix and comparison of experiment results are explained in Sect. 4. At the end, future scope and conclusion are stated in Sect. 5.

2 Related Work

The application of data mining in medicine allows doctors to predict various illnesses, understand the types of illnesses, particularly in neurology, and comprehend biomedicine. The main purpose of this work is to apply device learning and data mining methodologies to the ADNI dataset in order to identify the various tiers of AD. Many researchers nearly six distinctive tool studied and records mining algorithms has carried out to the ADNI dataset, inclusive of K-NN classifier, choice tree, GL model and deep neural net algorithm, so that it will define the five distinctive stages of the AD and to recognise the most distinctive characteristics via the variety of instances for EMCI and SMC commands, the performance of AD diploma kind may be except extra nice so that could teach the model for all commands with adequate and balanced records. Clinical exploration inside the past has demonstrated that AD pathology begins ten to fifteen years sooner than clear clinical indications and manifestations of intellectual problem begin to emerge in victims related to AD. Early guess of AD the utilisation of attainable biomarkers of early degree cerebrospinal liquid (CSF) may likewise be advantageous inside a clinical preliminary format. The right solution for victims with AD is clarified in [4]. AI methods were utilised to distinguish MRI pictures with Alzheimer's infection, addressed by Heba Elshatoury et al. [5]. The information utilised for the review is three-dimensional mind MRI examines. To prepare classifiers, in which the exactnesses were thought about, utilised directed learning calculations. For all cuts, all things considered, histograms are utilised. Then, at that point, chosen interesting cuts for additional examination dependent on the most noteworthy outcomes. The ADNI dataset was utilised and gave 69.5% precision. This current exploration's future degree incorporates further review is as yet being completed to improve the exactness of classifier and join more modalities by utilising positron outflow tomography pictures. During the flow research [6], we checked the adequacy of characterising admirably subject and AD patients utilising mind shape data. The discoveries showed 87.5% precision of characterisation, which was better than the exactness (81.5%), ordinarily utilised for conventional morphological adjustment appraisal. The Lebedev and different creators [7] clarify factors line clamor and post handling methodologies examined which are valuable clinical purposes. This exploration utilises MRI pictures and their estimations to examine the adequacy of RF classifier. Here, 225 HC and 185 AD are parted haphazardly into datasets for preparing and research, and the ADNI data set is utilised. Fan et al. [8], the model technique for the SVM is utilised to display, recognise and figure various cycles of AD to make a helper analysis of the illness dependent on MRI imaging information. Here, SVM model uses MRI information to create consequences of order. This shows that nearby is acceptable potential for the SVM philosophy to adjust to AD infection and can be utilised to test unsure examples. In the domain of image pattern recognition classification, deep learning algorithms have given state of art results [12–17]. Recent research is performed on detection and classification of melanoma skin cancer using deep learning and transfer learning algorithms [18–20]. During the overview, it has been seen that CNN calculation gives exact characterisation result than other AI calculations. Another perception is that CNN is more effective and less muddled while dealing with picture dataset.

3 System Architecture

In the proposed system, the CCNN classifier is used on MRI scans of AD patients. We changed the different tuning parameters in a such way that its produced the better testing accuracy than the existing. After acquiring the data, the MRI scans are rescaled in order to passed to layers of CNN. Each MRI scan is passed to convolutional as well as max pooling layer for feature extraction. The added extra not only convolutional but also max pooling layers helped to get the superior accuracy than existing systems. We have customised the CNN to achieve better results, first layer is the convolutional layer which we have used to convert the images into array of pixel values. For extraction of features from image, 3*3 filter is used and applied max pooling with 2×2 stride to reduce the dimensionality and to obtain the more exact features from the images along with Relu activation function, so that if there were any negative values in result that would be converted into zero. Likewise, we have used five hidden layers with the same process. Then we have flattened the results to feed into dense layer, and after that passed the results to fully connected layer and performed the softmax activation as we are performing multi-class classification, and finally, based

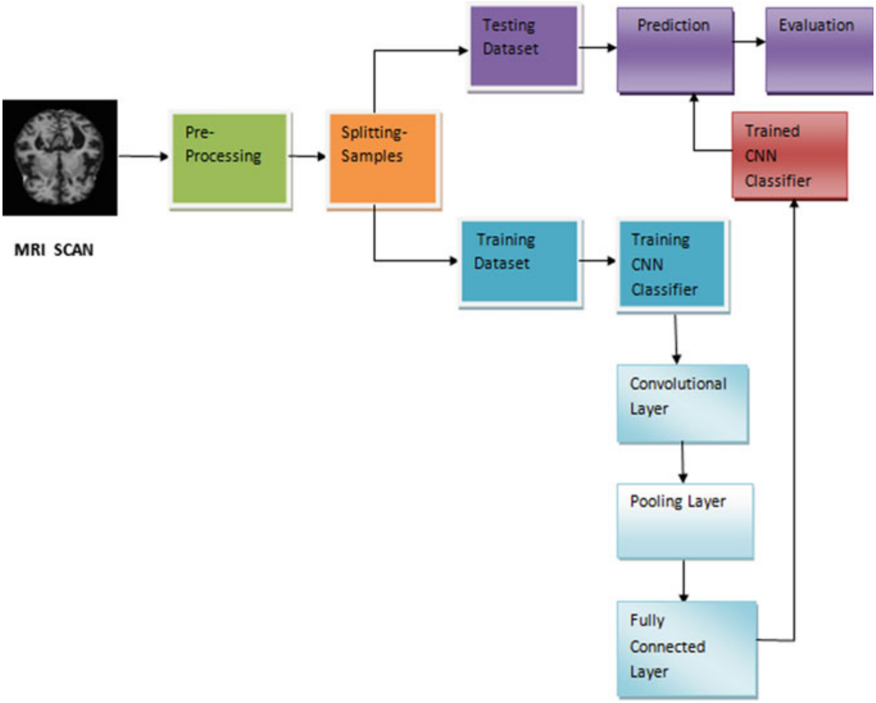


Fig. 1 System architecture

on the probabilities, we got the results. Figure 1 represents the CCNN architecture of proposed system.

3.1 *Magnetic Resonance Imaging (MRI)*

MRI is radiological method, which produces pictures of the body's anatomy and physiological functions. Images of the brain are created using MRI scanners by using powerful magnetic fields and radio waves. MRI scans are used as raw input image which can be used to produce activation maps that shows which parts of the brain are affected. It detects changes in blood oxygenation and flow that occurs in response in the neural activity.

3.2 Pre-processing

Image pre-processing is a technique for performing operations on images at the lowest level with the purpose of improving image features while eliminating undesirable distortions and boosting image information content without increasing image information content. The pre-processing stage includes image resizing, image transfer, and strength modification.

As images collected are different in size, we need to resize them before feeding to algorithm.

$$\text{Train\_datagen} = \text{ImageDataGenerator}(\text{rescale} = 1/255)$$

3.3 Splitting of Data

Data set is separated in two categories named training (utilised to train model) and testing (utilised to validate model performance).

$$[\text{setTrain}, \ \text{setTest}] = \text{partition}(\text{Dataset}, \ [0.8, \ 0.2], \ \text{randomized})$$

3.4 CCNN Learning Algorithm

A supervised CNN is a deep learning approach that accepts an image as input and uses learnable weights and biases to give priority to various objects in the picture, allowing them to be distinguished from one another. A CNN requires substantially less pre-processing than other classification algorithms. CNN model is represented in Fig. 2.

Different steps involved to learn convolutional neural network algorithm are:

Step 1: The system is given with a data collection including images and reference captions.

Step 2: The encoder is a convolutional neural network that extracts picture features pixel by pixel.

Step 3: The retrieved pixels are subjected to matrix factorisation. The matrix is of m x n.

Step 4: This matrix is subjected to maximum pooling, in which the highest value is chosen and then fixed back into the matrix.

Step 5: Every negative number is turned to zero as part of the normalisation process.

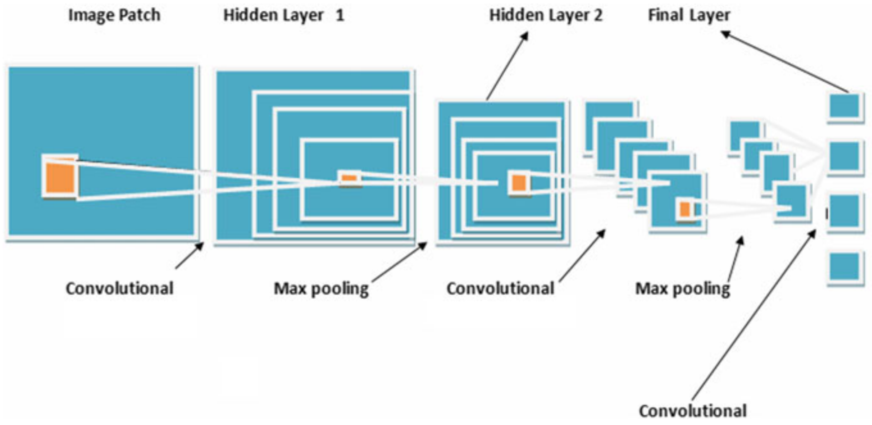


Fig. 2 Convolutional neural network

Step 6: To reduce a value to zero. Each value is filtered, and negative values are set to zero, resulting in corrected linear units.

Step 7: After determining the maximum probability, the hidden layers apply weights to the input values from the visible layers.

Mathematical model CNN learning algorithm:

Step 1: First we have to load the input images in a variable.

Step 2: Create a filter matrix and randomly initialise it, the filter is used to combine the images.

$$Z_1 = X * f$$

Step 3: On the result, use the sigmoid activation function.

$$A = \text{Sigmoid}(Z_1)$$

Step 4: Weight and baias matrices are defined randomly. Transform the values with the linear transformation.

$$Z_2 = W^{\mathrm{T}}.A + b$$

Step 5: On the data, use the sigmoid function. This will give the final output.

$$O = \text{Sigmoid}(Z_2)$$

3.5 Pseudo Code of CNN Learning Algorithm

```
Input:
        d: dataset, l: dataset true labels, t: target size
Output:
         Classified stage  of Alzheimer's Disease on MRI scan
         Let i be the MRI scan to be classified
        for i in dataset do ImageDataGenerator (rescale =1/255)
           ftrain, ftest, ltrain, ltest <—split feature set and labels
             into train subset and test subset
               S <—CNN (ftrain, ltrain)
           Let Conv2D be the convolutional and MPool2D be the max pooling layer
        for j in layers do
                tf.keras.layers.Conv2D(activation='relu', input_shape=())
                  tf.keras.layers.MPool2D(2, 2)
                Let DIR be the path joining version for other paths
          for k in categories do
                DIR = main_path+'category'
                  dataset_length.append(category_length)
                   Result <—- evaluate (DIR,ltest, S)
     Return result;
```

There are four categories of images [mild, moderate, none and very mild] after applying CNN it will provide the predicted class of the image.

4 Dataset and Experiment Results

Between 2004 and 2017, the ADNI collected clinical, genetic and imaging data from approximately 1550 participants, such as AD patients, elderly controls and MCI subjects, across three multi-year cohorts (ADNI1, ADNI GO and ADNI2).This understanding aids in the development of better the clinical studies for the prevention and remedy of AD. There are four categories named as mild, moderate, none and very mild with total 6400 MRI scans. Dataset is downloading from the Kaggle and some samples of dataset are represented in Fig. 3.

These are MRI scans from each category of an AD patients. Data pre-processing is the phase in any machine learning process when the data is altered or encoded to make it easier for the machine to parse it. To put it another way, the algorithm can now readily comprehend the characteristics of the data. The dataset is partitioned into two categories: training set (utilised for developing model) and test set. Test set is a sample of data that is used to assess the created model objectively. From the original dataset, image data augmentation provides new training samples. The model is created by making random changes to the layout of the photos while keeping

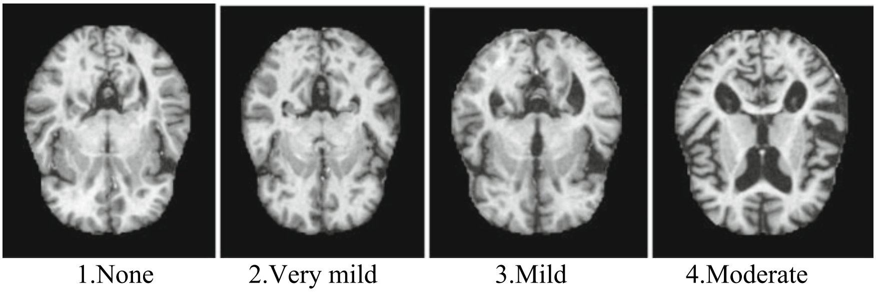


Fig. 3 Four categories of Alzheimer dataset

the data's class labels. The image data generator method in Keras is used for data augmentation in the system that we have implemented. As an argument, the method expects an array of images. To achieve augmentation, it applies a series of random alterations to each image in the array. Finally, the calling method receives the array of augmented photos.

In Fig. 4, in graph epochs are shown on X-axis and loss is shown on Y-axis. With each epochs, loss is decreasing, hence leading to better accuracy.

Figure 5 in Graph 2 has epochs on X-axis and accuracy on Y-axis. As accuracy and loss are inversely proportional, epochs and accuracy increases while the loss decreases.

The above matrix is utilised to calculate classification performance of a model based on test data with true values which is represented in Fig. 6. The formula is (TP + TN)/Total to calculate the accuracy. The true positive rate defines how many times the classifier predicts result as yes when it is actually yes. Table 1 helps to understand the existing accuracies of algorithms, and it can be stated that random

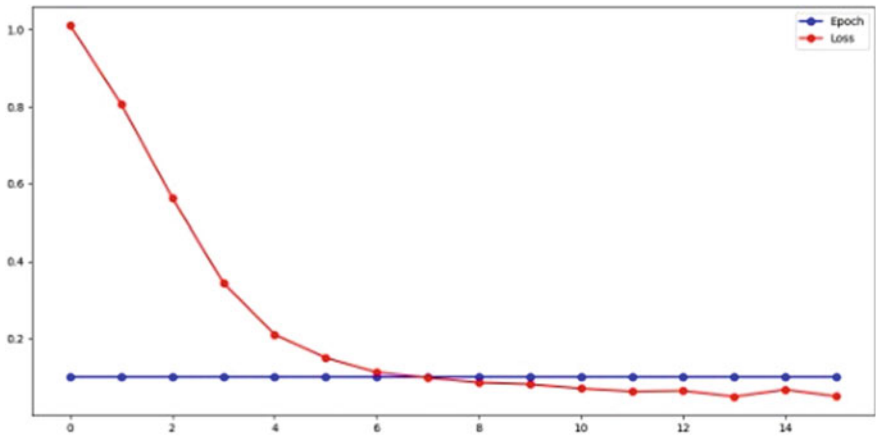


Fig. 4 Graph 1: loss graph

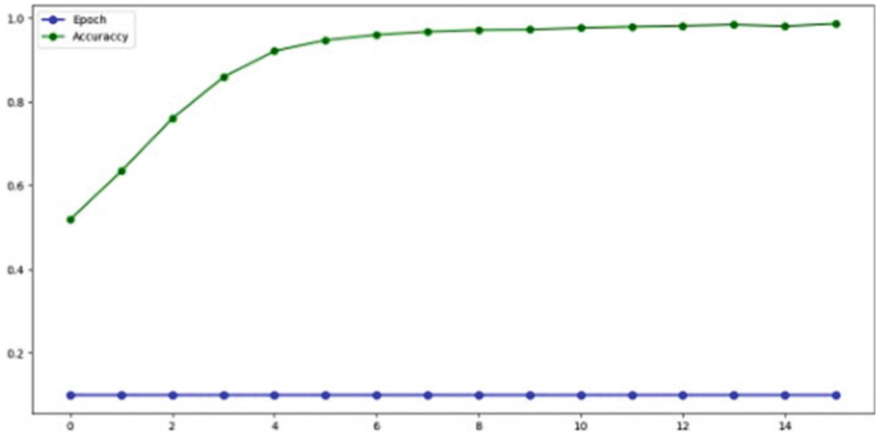


Fig. 5 Graph 2: accuracy graph

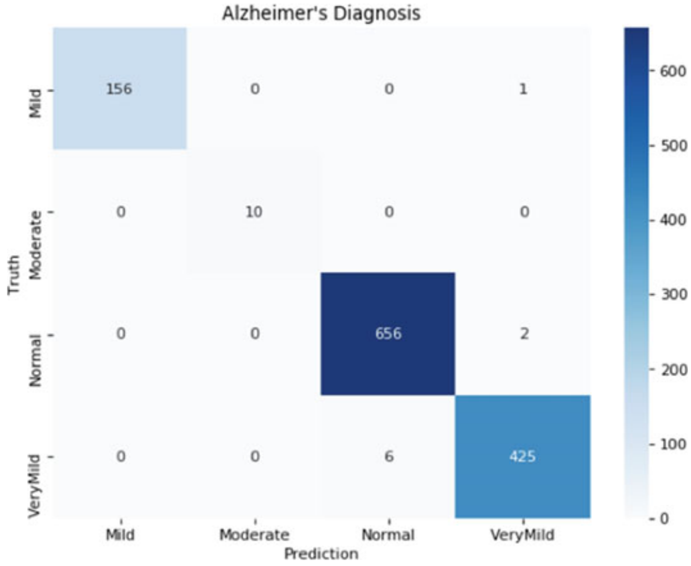


Fig. 6 Confusion matrix

forest and support vector machines are the two algorithms that give accuracy closer to CNN algorithm.

Table 1 Result comparison CCNN with existing algorithms

Year and Ref.	Method	Accuracy (%)
(2019) [8]	Convolutional neural network	85
(2019) [9]	K-NN, decision tree	88.24
(2017) [4]	Naïve Bayes, sequential minimal opt	82.19
(2019) [10]	All pairs technique	86.66
(2019) [5]	Support vector machine	69.5
(2019) [11]	Random forest	85.5
Proposed system	Customised CNN	97.21

4.1 Different Challenges to AD Detection and Classification

Though the ADNI has provided dataset of MRI scans of AD patients, the dataset is not sufficient and adequate. There are four categories and data is not equally distributed in those categories, resulting unstable dataset as well as poor performance of model.

Only MRI scans are not enough for precise classification of AD stage as there are other factors and symptoms causing the advancement of disease. Currently, we have used only MRI scans to classify AD, but for more accurate results, we can use various sensors to observe and keep track of symptoms of the patient for prediction and classification. This method will require monitoring routine and biological changes in the body for longer period of time. The disadvantage is that it requires huge data storage and more time for analysis.

5 Conclusion

This paper explains the proposed system for AD classification based on MRI scans of ADNI dataset using customised convolutional neural network. The survey we conducted for the disease classification led us to the observation that CNN performs better than the other algorithms. One of the challenges is acquiring dataset as there is not enough data for processing. The machine learning algorithms classify with better accuracy if provided larger dataset. As stated in the future scope, there are few other ways to collect more data with different features for advancement of classification rate. Here, we can conclude that the proposed CNN system explained in this paper is more efficient than other existing algorithms with 97.21% accuracy.

References

1. Prince M et al (2014) World Alzheimer report
2. Blennow K (2007) Alzheimer's disease. Lancet 368(9533):387–403
3. Yiannopoulou KG et al (2013) Current and future treatments for Alzheimer's disease. Ther Adv Neurol Disord 6(1):19–33
4. Asif Hassan S et al (2017) A machine learning model to predict the onset of Alzheimer disease using potential cerebrospinal fluid (CSF) biomarkers. Int J Adv Comp Sci Appl 8(12)
5. Elshatoury H et al (2019) Volumetric histogram-based Alzheimer's disease using support vector machine. J Alzheimer's Dis
6. Hiroki F et al (2018) Detection of Alzheimer's disease with shape analysis of MRI images. In: Joint 10th international conference on soft computing and intelligent systems (SCIS) and 19th international symposium on advanced intelligent systems (ISIS), pp 1031–1034
7. Lebedev AV et al (2014) Random forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness. Neuro Image Clin 6:115–125
8. Fan Z et al (2019) Classification of Alzheimer's disease based on brain MRI and machine learning. Neural Comput Appl 32:1927–1936
9. Raza M et al (2019) Diagnosis and monitoring of Alzheimer's patients using classical and deep learning techniques. Expert Syst Appl 136:353–364
10. Zhang Y et al (2012) An MR brain image classifier via principal component analysis and kernel support vector machine. School Inf Sci Eng 130
11. Xu L et al (2019) k-Skip-n-Gram-RF: a random forest based method for Alzheimer's disease protein identification. Front Genet 10
12. LeCun Y et al (1989) Handwritten digit recognition with a back-propagation network. Adv Neural Inf Process Syst NIPS 2:396–404
13. Mane D, Kulkarni U (2017) A survey on supervised convolutional neural network and its major applications. Int J Rough Sets Data Anal 4:71–82
14. Mane D, Kulkarni U (2018) Visualizing and understanding customized convolutional neural network for recognition of handwritten Marathi numerals. Proc Comp Sci 132:1123–1137
15. Mane DT et al (2021) Vehicle-type classification using customized fuzzy convolutional neural network. In: Data engineering and intelligent computing. Advances in intelligent systems and computing, Springer, Singapore, Vol 1, pp 419–429
16. Mane DT et al (2021) Handwritten Marathi numeral recognition using stacked ensemble neural network. Int J Inf Technol 13:1993–1999. https://doi.org/10.1007/s41870-021-00723-w
17. Shinde S et al (2021) Deep learning-based medical image analysis using transfer learning. In: A computational perspective in healthcare. studies in computational intelligence, Springer, Singapore, Vol 932, pp 19–42
18. Patil R (2021) Machine learning approach for malignant melanoma classification. Int J Sci Technol Eng Manage A VTU Publ 3(1):40–46
19. Patil R, Bellary S (2021) Transfer learning based system for melanoma type detection. Revue d'Intelligence Artificielle 35(2):123–130
20. Patil R, Bellary S (2020) Machine learning approach in melanoma cancer stage detection. J King Saud Univ Comp Inf Sci 1–9. https://doi.org/10.1016/j.jksuci.2020.09.002

A Deep Dive Comparison of Cache Replacement Strategies: The Quality of Experience Influencer

C. Victoria Priscilla and A. R. Charulatha

Abstract Customer's satisfaction has been a prime concern to most companies in designing and refining their products and their services. The background of rising of the user's quality of experience (henceforth QoE) concept has given rise to new services. These services have majorly contributed in increasing the profits of the Internet service providers; however, their revenue growth is strongly dependent on the perception and feedback of the end users. Content-centric networking (CCN) is a rising new architecture which paves way for a looming Internet. The efficiency of this architecture is evaluated based on the caching strategies that is being implemented. One approach in refining the user's quality of experience (QoE) is caching. The performances of cache replacement strategies in CCN are evaluated and a detailed comparison is done with relevant caching strategies on different cache sizes. The performances in terms of the metrics are analysed.

Keywords Cache replacement · CCN · QoE · LRU · LFU · Random cache · Semantic · Cache

1 Introduction

The swift development in the Internet and the rise of Internet applications has not only increased the traffic in business network, but also amplified the user's demand for proper bandwidth and faster delivery from the service providers [1]. Meeting up with user's expectation and their requirements with regard to reliable and access to anytime, anywhere service, quality and cost-effective communication bandwidth are

C. Victoria Priscilla (✉)
Department of Computer Science, Shrimathi Devkunvar Nanalal Bhatt Vaishnav College for Women (Autonomous), Chrompet, Chennai 600044, India
e-mail: aprofvictoria@gmail.com

A. R. Charulatha
Stella Maris College (A), University of Madras, Chennai 600086, India
e-mail: charulatha@stellamariscollege.edu.in

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_34

the areas that needs to be addressed. These key aspects of scalability, mobility and security have caused major impact in the user's quality of experience.

Contrasting to IP-based which is a host-oriented architecture content-centric networking (henceforth CCN) accentuates the data by making it directly addressed and routed. The network implements a unified naming scheme for the content. The source and destination communicated based on the content names that are mapped to IP [2]. CCN majorly follows in-network caching where several copies of the same data are stored in the network. Any node present in the network path which has the content can serve the future requests in their path, thereby reducing the network load and bottleneck in the network which improves the QoE. This architecture of CCN has enabled to handle the massive rise in Internet applications and consumer's demand for high bandwidth utilization and minimal delay in delivery [3].

2 Background Study

2.1 The CCN Framework

In CCN, two transmission packets are used, they are: Interest packet and data packet (hereafter DP). A user creates the requested content as interest packet which holds the name of the content and it is forwarded through router nodes. When an interest packet arrives the node, which is the content provider, a DP having the requested content is directed to the requested end user. The DP contains the name of the content, the content and content provider's signature along the same requested path. The sender's or the receiver's address is not mentioned in the transmission packets in this entire communication of request and response [4].

In CCN architecture, broadcasting is accomplished mainly with the use of three tables. The content store caches the DPs to respond to the interest packets which will arrive in future. The pending interest table stores the interest packets and the node information to which the request has been forwarded but have not yet received the content. This helps CCN to achieve the collection and storing of information at one place, avoid requesting for the redundant interest packets often and avoid network resources wastage. The purpose of forwarding information base is to store the neighbouring node or next-hop node information that helps to reach the content provider and also is used to respond for the interest packets [5].

2.2 User's Quality of Experience

An end user in any network prefers to have access to various applications and services without any limitations like time, device type, mobility, location and technology. This has now become the primary goal of any Internet service provider to sustain their customer base. This has led to integrating the existing and new technologies in wireless communication which aims to fulfil the end user's expectations [6]. This is referred to as user's quality of experience (QoE).

The major aim of network provider in QoE management may be to optimize the consumer QoE at the same time making efficient utilization of network resources and sustaining a satisfied customer base. To achieve this, it is first of all necessary to understand and identify the subjective and objective factors which affects the QoE for the application from the consumer's point of view and also its level of impact. With regard to wireless networks, a user's experience is assessed based on the following factors like performance, accessibility, reliability, latency, jitter, network quality, content effectiveness, usage history effectiveness, server reliability and availability. Out of many device-based factors and network-based factors affecting the user's QoE, network factors are the challenges that need to be addressed for sustaining the customer base by the network service providers.

The paper is organized in the following manner: In Sect. 2, the need for caching, caching strategies and cache replacement strategies is discussed. The simulation environment of the evaluation scenario, the comparative study results and analyses the efficiency of each strategy are discussed in Sect. 3. The summary of findings and conclusion is presented in Sect. 4.

3 Motivation

Among other factors influencing the user's QoE, the caching strategy has more impact in improving the performance of content distribution in CCN and hence has gathered lot of interest among the researchers in the recent years.

A significant feature of CCN is that every router node in the network has the ability to integrate a data cache, or content store. To improve the QoE most contents that are popular can be cached in the network nodes that are nearer to the edge, resulting in swift response to the consumer's request [7, 8].

Owing to the space constraints in CS, it seems to be hard to attain unrestricted content caching. Hence, when the upper limit is reached for a cache space usage or an unpopular content is cached, the relatively less popular contents have to be removed or replaced to give up the space for popular content. On what basis the cached content is deleted or updated from the CS is decided by the cache replacement strategy, and it is now focal point of CCN cache performance improvement research. The flowchart representation of the caching and cache strategy is depicted in Fig. 1.

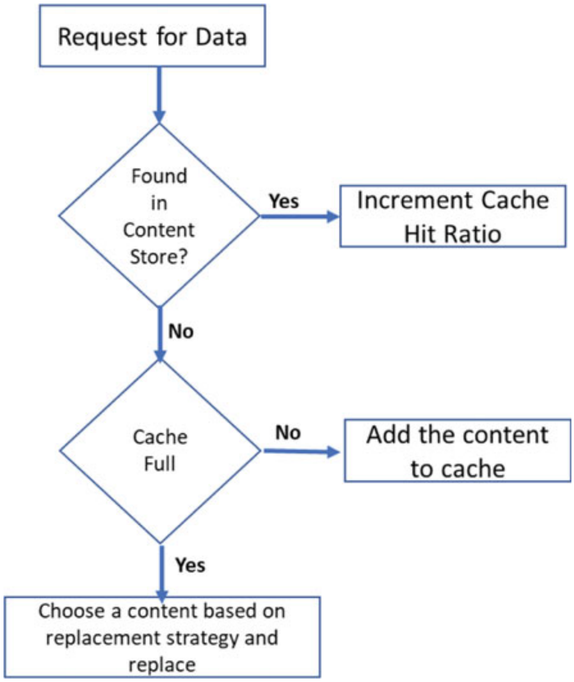


Fig. 1 Flowchart for caching and cache replacement strategy

3.1 *Caching Strategies*

The main caching strategies leave copy everywhere (LCE) [9], leave copy down (LCD) [10], MAx-Gain in-network caching (MAGIC) [11] and probabilistic in-network caching (ProbCache) [12] are the caching strategies which are considered for evaluation. The main disadvantages of many of these caching strategies are the redundancy of caching data and also the rate at which cache replacements occur is higher. Thus, the focus is to provide an efficient caching strategy, which reduces the redundancy in CS and improves the cache hit. However, more the priority is given to caching strategies to have better cache hit ratio has led to overloading in the central nodes and thereby leads to frequent cache replacements which thereby reduces the subsequent utilization of caching resource. This has led to concentrate on choosing the appropriate cache replacement strategy which also has a key role in improving the cache hit ratio.

3.2 *Cache Replacement Strategies*

The data and its information which are cached cannot be stored in the CS for longer time due to the limited CS size and the catalogue nature of caches. So, it becomes

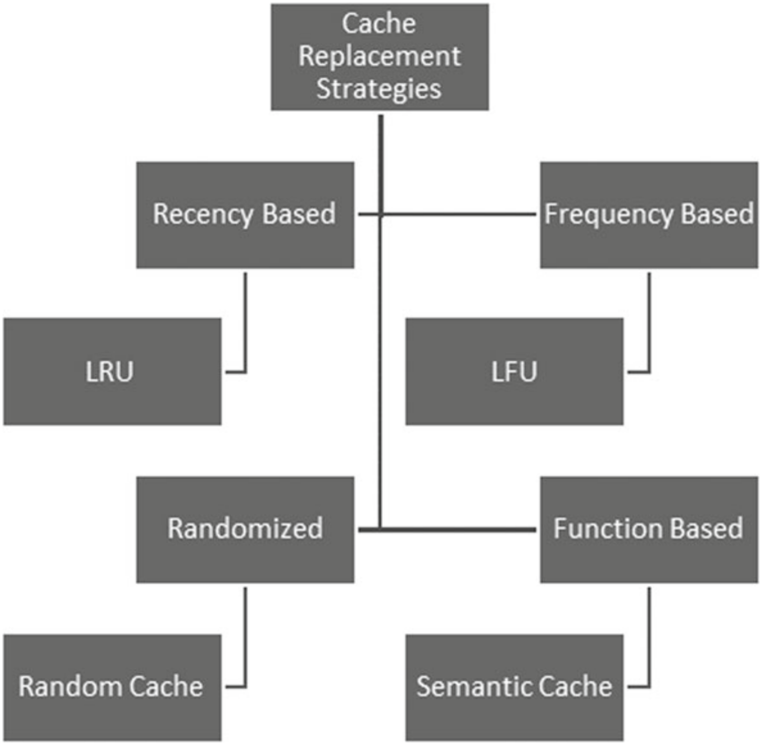


Fig. 2 Cache replacement strategies

necessary to replace the content in CS whenever the CS is full and when a new request arrives. Several replacement policies have been considered and implemented in CCN. In this paper, we shall analyse the LRU, LFU, random cache replacement and semantic cache replacement strategies. The methodology of the cache replacement strategy is represented in Fig. 2.

A. **Least Recently Used**

Least recently used (LRU) is a replacement strategy, in which when a content has been requested, LRU assumes that it will be requested again in near future, so the content which is arrived recently is replaced with the content in the content store that has not been demanded or requested for a long period of time. LRU is a replacement policy that is widely implemented as it is easier compared to other strategies, but the actual drawback is that LRU does not take into consideration the popularity of content. A content which is very much popular is possible to be replaced by the content just arrived, but less popular content [13].

B. **Least Frequently Used**

This strategy decides the popularity of contents based on the number of times that content is requested by a user. When the CS is fully exhausted and a new content request arrives, the content that is requested least number of times will be considered for replacement. The main disadvantage of this LFU strategy is that it holds the content which is no longer popular and still allows to be in the CS occupying the space which results in unnecessary space utilization.

C. **Random Cache Replacement**

Random (Rand) cache replacement strategy uses a uniform distribution. The contents chosen for replacement are uniform after the CS space is exhausted [14] in the nodes. The main drawback of this strategy is that many researches have proved that this strategy replaces the most popular content and retains the non-popular contents.

D. **Semantic Cache Replacement**

Semantic replacement is a replacement strategy that follows a defined syntax and rules. In this strategy, the tags, fields and conditions of the replacement strategies will be a described by cache administrator, this is called the grammar nature of the strategy. The strategy of setting a semantic description of data that is cached in the content store allows the use of value functions that utilizes the semantic notions for cache replacement. The disadvantage of this strategy is the rigid set of rules defined. As these rules are inflexible, a complete redesign has to be done if the most popular content is replaced by mistake [14].

4 Results and Discussions

A discrete-event simulation tool is implemented and available at [15] is used to compare the replacement strategies in similar simulation environment.

The simulation experiments are performed on the four cache replacement strategies by varying the parameters and caching strategies as listed in Table 1. Cache hit ratio, diversity and stretch are used to analyse the cache replacement strategy that performs best with every cache strategy. By varying the cache size and caching strategies the simulation environment is experimented with Facebook social media data set.

Table 1 Simulation setup

Replacement strategies	LRU, LFU, random cache, semantic cache
Caching strategies	LCE, LCD, ProbCache, MAGIC
Metrics for evaluation	Cache hit, diversity and stretch
Cache size	Up to 10^5
Topology	Abilene

Table 2 (**a**) Cache hit ratio for LRU and LFU. (**b**) Cache hit ratio for rand cache and semantic cache

(a)

Cache size	LRU				LFU			
	LCE	LCD	Magic	ProbCache	LCE	LCD	Magic	ProbCache
10	0.0625	0.0769	0.0769	0.0	0.0714	0.0357	0.0625	0.125
100	0.0714	0.0625	0.0833	0.0714	0.0714	0.0677	0.0714	0.0
1000	0.0588	0.556	0.0625	0.0833	0.0588	0.0714	0.0667	0.083
10,000	0.0667	0.0833	0.0526	0.0909	0.0833	0.0625	0.0714	0.0
100,000	0.0526	0.1	0.0667	0.0	0.1111	0.0435	0.0833	0.0714

(b)

Cache size	Rand cache				Semantic cache			
	LCE	LCD	Magic	ProbCache	LCE	LCD	Magic	ProbCache
10	0.05	0.0625	0.0667	0.0625	0.0667	0.0667	0.0588	0.0
100	0.0526	0.0667	0.0556	0.1111	0.0714	0.0714	0.0588	0.0
1000	0.0566	0.0526	0.0714	0.0435	0.0588	0.0526	0.0769	0.0625
10,000	0.0588	0.0714	0.0667	0.0909	0.0526	0.0909	0.1	0.0
100,000	0.0667	0.05	0.0625	0.0526	0.0667	0.05	0.0714	0.0

Cache Hit Ratio: Higher the cache hit ratio shows reduced load on server and maximum efficiency, ensuing in less delay in transmission. With respect to Table 2a, b it can be understood that least recently used produces better results with LCD and ProbCache caching strategies for high cache sizes and moderate results with LCE. Random cache works efficiently with ProbCache caching as cache size increases. Semantic cache shows better results with LCD and magic and does not favour ProbCache. LFU works efficiently with LCE, LCD and MAGIC as cache size increases (Fig. 3). Thus, least recently used and random cache techniques seem to be efficient for achieving cache hit ratio.

Diversity

The diversity value ranges between [1/|N|, 1]. If the value inclines to 1/|N|, then it shows that caches contain the redundant content all over the network; else, if the value of diversity inclines to 1, then the CS caches unique data in every cache [16]. With reference to Table 3a, b LRU works efficiently with LCD irrespective of all cache sizes also works fine with ProbCache for smaller cache sizes and has lot of redundant values stored when used with LCE or MAGIC. Random cache works efficiently with LCD as cache size increases. Semantic cache works fine for LCD for all cache sizes. LFU works efficiently with LCD for all cache sizes and Probcache for smaller cache sizes (Fig. 4). Thus, we may conclude LFU and LRU seems to be effective with respect to the diversity metric.

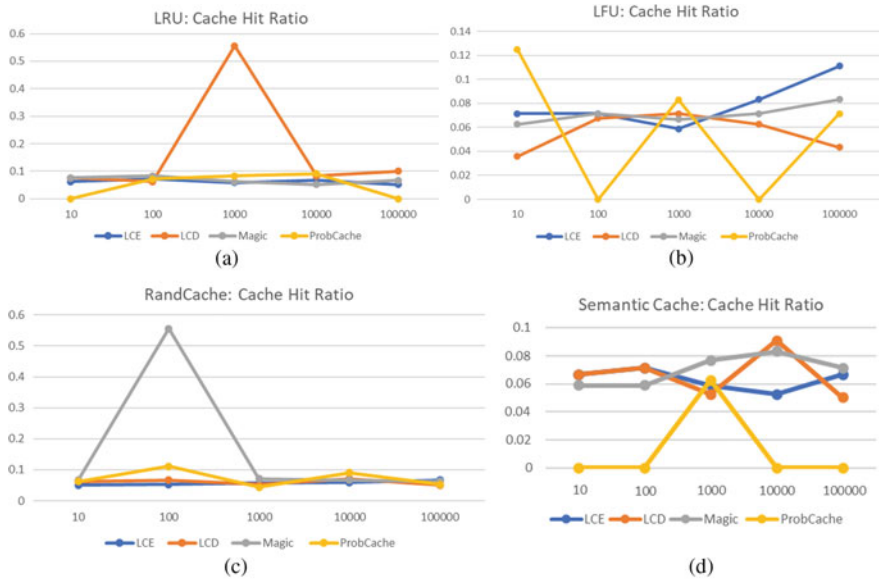


Fig. 3 Cache hit ratio with respect to **a** LRU **b** LFU **c** rand cache **d** semantic cache

Table 3 (**a**) Diversity for LRU and LFU. (**b**) Diversity for rand cache and semantic cache

Cache size	LRU				LFU			
	LCE	LCD	Magic	ProbCache	LCE	LCD	Magic	ProbCache
10	0.2667	0.8	0.3333	0.6667	0.3077	0.8	0.2667	0.6667
100	0.3077	0.8	0.3636	0.4444	0.3077	0.8	0.3077	0.6667
1000	0.25	0.8	0.2667	0.5714	0.25	0.8	0.2857	0.4444
10,000	0.2857	0.8	0.2222	0.5714	0.3636	0.8	0.3077	0.4
100,000	0.2222	0.8	0.2857	0.3636	0.5	0.8	0.3636	0.5
Cache size	Rand cache				Semantic cache			
	LCE	LCD	Magic	ProbCache	LCE	LCD	Magic	ProbCache
10	0.2105	0.8	0.2857	0.4	0.2857	0.08	0.25	0.4444
100	0.2222	0.8	0.2353	0.5714	0.3077	1.0	0.25	0.3333
1000	0.2353	0.8	0.3077	0.3333	0.25	0.8	0.3333	0.4444
10,000	0.25	1.0	0.2857	0.4444	0.2222	0.8	0.3636	0.4444
100,000	0.2857	0.8	0.2667	0.5	0.2857	0.8	0.3077	0.4444

Stretch

While the stretch value moves towards 0, it shows that the requests are responded by the neighbouring nodes of the consumers. With reference to Table 4a, b, it may be concluded that LRU is not satisfying this request as all the values are nearing

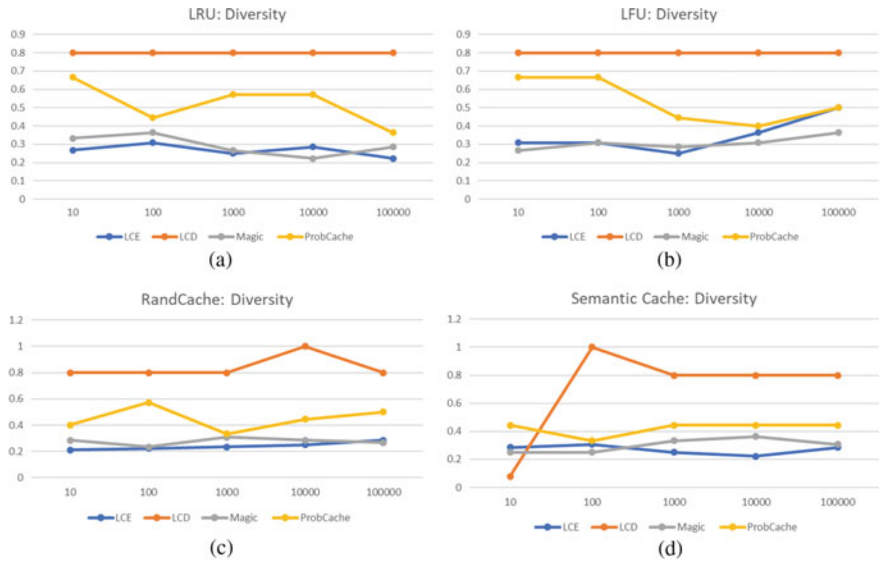


Fig. 4 Diversity with respect to **a** LRU **b** LFU **c** rand cache **d** semantic cache

Table 4 (**a**) Stretch for LRU and LFU. (**b**) Stretch for rand cache and semantic cache

Cache size	LRU				LFU			
	LCE	LCD	Magic	ProbCache	LCE	LCD	Magic	ProbCache
10	0.9167	1.0	1.0	1.0	0.9	1.0	1.0	1.0
100	1.0	1.0	0.875	1.0	1.0	1.0	0.9	1.0
1000	1.0	1.0	0.8462	1.0	0.9231	1.0	1.0	1.0
10,000	1.0	1.0	1.0	1.0	1.0	1.0	0.8182	1.0
100,000	1.0	1.0	0.9091	1.0	1.0	1.0	0.875	1.0
Cache size	Rand cache				Semantic cache			
	LCE	LCD	Magic	ProbCache	LCE	LCD	Magic	ProbCache
10	1.0	1.0	0.8	1.0	0.8333	1.0	1.0	1.0
100	0.9333	1.0	1.0	0.8	1.0	1.0	1.0	1.0
1000	0.8667	1.0	1.0	0.9474	1.0	1.0	1.0	1.0
10,000	0.9231	1.0	1.0	0.8571	1.0	1.0	0.875	1.0
100,000	0.9091	1.0	0.9167	1.0	1.0	1.0	0.9	1.0

1. Rand cache works fine with LCE as cache size increases. Semantic cache works efficiently with LCE for smaller cache sizes and MAGIC for large cache sizes. LFU works efficiently with MAGIC for larger cache sizes (Fig. 5). With respect to the metric stretch, we may draw a conclusion that semantic cache seems to have better stretch value with LCE and MAGIC and LFU seems to be better.

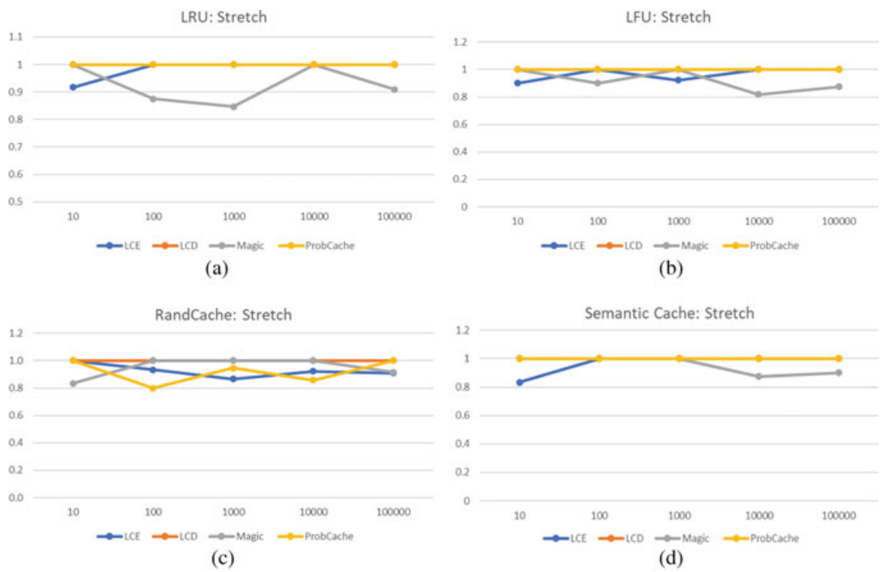


Fig. 5 Stretch with respect to **a** LRU **b** LFU **c** rand cache **d** semantic cache

5 Conclusion

Comparison of various cache replacement strategies in CCN by analysing advantages and disadvantages of each replacement strategy with regard to varied caching strategies and different cache sizes is implemented. The analysis therefore concludes that LRU replacement strategy has higher performance with regard to less consumption of bandwidth and better throughput since the cache hit ratio is better compared to other strategies. When considering the efficiency of LFU, it seems to satisfy cache hit ratio, diversity and Stretch with a limitation on cache size. Random cache seems to be efficient as cache size increase. A combination of replacement strategies would be effective for all cache sizes and caching strategies.

References

1. Li F, Lam KY, Wang L, Na Z, Liu X, Pan Q (2018) Caching efficiency enhancement at wireless edges with concerns on user's quality of experience. Wireless Commun Mob Comput
2. Ji Y, Zhang X, Liu W, Zhang G (2021) Replacement based content popularity and cache gain for 6G content-centric network. Phys Commun 44:101238
3. Kumar S, Tiwari R (2020) Optimized content centric networking for future internet: dynamic popularity windowbased caching scheme. Comput Netw 179:107434
4. Wissingh B, Wood CA, Afanasyev A, Zhang L, Oran D, Tschudin CF (2020) Information-centric networking (ICN): content-centric networking (CCNx) and named data networking (NDN) terminology. RFC 8793:1–17

5. Mastorakis S, Afanasyev A, Moiseenko I, Zhang L (2016) ndnSIM 2: an updated NDN simulator for NS-3. NDN, Technical Report NDN-0028, Revision 2
6. Yang M, Wang S, Calheiros RN, Yang F (2018) Survey on QoE assessment approach for network service. Ieee Access 6:48374–48390
7. Zhang J, Weiliang XIE, Fengyi YANG, Zhouyun WU, Liang XIE (2016) Mobile edge computing and application in traffic offloading. Telecommun Sci 32(7):132
8. Zhang M, Luo H, Zhang H (2015) A survey of caching mechanisms in information-centric networking. IEEE Commun Surv Tutorials 17(3):1473–1499
9. Jacobson V, Smetters DK, Thornton JD, Plass MF, Briggs NH, Braynard RL (2009) Networking named content. In: Proceedings of the 5th international conference on Emerging networking experiments and technologies, pp 1–12
10. Zhang G, Li Y, Lin T (2013) Caching in information centric networking: a survey. Comput Netw 57(16):3128–3141
11. Ren J, Qi W, Westphal C, Wang J, Lu K, Liu S, Wang S (2014) Magic: a distributed max-gain in-network caching strategy in information-centric networks. In: 2014 IEEE conference on computer communications workshops (INFOCOM WKSHPS), IEEE, pp 470–475
12. Psaras I, Chai WK, Pavlou G (2012) Probabilistic in-network caching for information-centric networks. In: Proceedings of the second edition of the ICN workshop on Information-centric networking, pp 55–60
13. Bilal M, Kang SG (2014) Time aware least recent used (TLRU) cache management policy in ICN. In: 16th International Conference on Advanced Communication Technology, IEEE, pp 528–532
14. Abdullahi I, Arif S, Hassan S (2015) Survey on caching approaches in information centric networking. J Netw Comput Appl 56:48–59
15. Cesar B (2014) Social CCN SIM is a CCN simulator
16. Priscilla CV, Charulatha AR (2019) A comparative study on caching strategies in content centric networking for mobile networks. In: 2019 11th international conference on advanced computing (ICoAC), IEEE, pp 122–128

Unusual Event Detection from Surveillance

G. Tamilmani, S. Uma, and S. Sivakumari

Abstract Detecting the unusual events is being difficult in intelligent video research. Though the technology is developed enormously in recent years, there is still lack of some concepts which makes easy and reduce manpower to do the things those require huge amount of manpower and time. Regarding this issue, there is a huge challenge in detecting the unusual events, if a person drives on a road it is normal event, but if the same person drives on pedestrian road it is not a normal event, there are many aspects to consider it as abnormal event, when it comes to normal surveillance it is impossible to understand which is abnormal and normal, only the video can be recorded but cannot classify the event, thus unusual event detection from surveillance is designed to classify the normal and abnormal events through anomaly detection process. Convolutional layers have showed potential in new applications of object detection and recognition mostly in images. Convolutional neural networks, on the other hand, are supervised and required labels to learn signals. Spatiotemporal recurrent neural network is used to learn the model by combining two primary components: spatial feature representation and temporal feature evolution. According to practical results on the avenue subway, the detection correctness of our system is equivalent to state-of-the-art methods at a significant speed of up to 45 frames per second.

Keywords Spatial features · Temporal features · Convolutional neural networks · Spatiotemporal recurrent network

G. Tamilmani (✉)
Department of Computer Science and Engineering, Veltech Rangarajan Dr. Saguthala R&D Institute of Science & Technology Avadi, Chennai, India
e-mail: tamilmanig@veltech.edu.in

S. Uma
Department of Computer Science and Engineering, Rajalakshmi Institute of Technology, Chennai, India

S. Sivakumari
Department of Computer Science and Engineering, Avinashilingam Institute of Home Science and Higher Education for Women, Coimbatore, India

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_35

1 Introduction

With the ever-increasing amount of video data, there is a growing requirement for not just object detection and methods, but also for spotting strange, exciting occurrences of odd things in a suspicious manner inside the huge structure of conventional data. Detecting abnormalities in movies is critical in a variety of applications, from autonomous internal control to video surveillance. Significant incidents, like as surveillance film, are exceedingly unlikely to occur on a daily basis. Manually detecting events or irregularities can be a very meticulous operation that always necessitates more staff than is often available. The challenge of detecting anomalies prompted the need for automated detection and segmentation of sequences of interest. Despite the fact that current technology necessitates a large number of configuration attempts on each video stream prior to the deployment of the video analysis process, events are still subjected to some predefined examining, making it difficult to generalize the detection model to different surveillance scenes. Video data is difficult to represent and model due to its high dimensionality, noise, and enormous number of events and interactions.

Abnormalities also are high in context, for instance, running into a shop is abnormal, but running in Public Park is normal. Though, the definition of abnormality is often ambiguous and roughly defined. One might imagine that walking around subway platform is normal but another might imagine that it should be flagged as an abnormality because of suspicion. These challenges made it difficult machine learning method to detect video patterns of abnormalities in real-world applications. Though current detection methods are only applicable to labeled video footage with clearly stated events of interest and no scenes that are highly obstructive, such as crowded scenes, current detection methods are only applicable to labeled video footage with clearly stated events of interest and no scenes that are highly obstructive, such as crowded scenes. Furthermore, the cost of labeling each and every type of event is enormous. Even so, it cannot ensure that it will cover all past and future events. The recorded video footage is unlikely to be long enough to cover all types of activity, particularly unusual or never-before-seen ones.

The effectiveness of a new effort to detect abnormalities by approaching the task as a binary classification issue has been established, but the practicality of such methods is limited because film of aberrant actions is difficult to gather due to its peculiarity. As a result, many researchers have focused on models including spatiotemporal characteristics, dictionary learning, and auto-encoders that can be implemented with little to no supervision. These methods, unlike supervised methods, merely require unlabeled video footage with few or no aberrant events, which is simple to get in real-world applications. The following part will go through these approaches in detail, as well as their limitations. This research proposes a unique framework for representing a deep learning approach, and video data is automatically retrieved from a long video using a set of general attributes. Unsupervised video frame analysis was carried out using a deep neural network made up of a stack of convolutional auto-encoders, which carried off spatial characteristics in the input that gathered to create the video representation.

Fig. 1 Unusual event occurrences

The representation is then input into a stack of convolutional temporal auto-encoders, which are used to comprehend the regular temporal patterns. The method of presentation is domain-independent; it does not necessitate any additional human labor and may be applied to a variety of settings. We use a real-world dataset to demonstrate the accuracy of the presented method, and we show that our method regularly outperforms similar methods while keeping a quick running time. Figure 1 depicts the occurrences of unusual event such as traffic violations, robberies, and natural cause.

2 Related Works

Anomaly detection has been the biggest challenge in this era [1]; technology has been grown to find the abnormal events from having manpower physically observing to the stage; a system can completely record the whole situation, and at this point, there lies a problem by recording the whole situation whole time; there is enormous amount of data, to detect some event in this huge data; there are only two ways either having manpower to continuously observe or checking the recorded situations to find the unusual events, either way it costs extreme manpower or system power. The majority of irregular patterns is unknown in advance, as this will necessitate forecasting all of the techniques that could go wrong. However, learning a model for everything that is abnormal or irregular is impossible, but how can you locate an anomaly if you do not know what to look for? As a result, instead of focusing on detecting what is exceptional, it is possible to concentrate on a scenario where the training data only contains common visual patterns. Researchers in this field have adopted a widely held perspective in which they first learn the usual patterns from training videos, and then look for abnormalities as events deviate from the usual patterns [2].

The majority of abnormality detection research is based on extracting local behaviors from films and using them to build a normalcy model. In the process of grouping visual words, traditional BOV techniques destroy any compositional information by grouping comparable volumes. It is also necessary to check the number of clusters ahead of time, which can only be accomplished during testing through try and error. Despite their simplicity, recently presented techniques and methodologies have several limitations.

Despite the fact that 3D ConvNet excels at distinguishing anomalies from normal events and activities, it is currently impossible to employ in real-world applications due to a shortage of video segments featuring abnormal actions [3]. While convolution and pooling techniques are only used in a limited number of situations to performed spatially in the convolutional auto-encoder stated by the stated network takes multiple frames as input, and in the case of 2D convolutions, temporal information is completely crumbled after the first convolution layer. Aside from the fact that convolutional LSTM layers are memory intensive, training will have to be done in small batches, resulting in long training and testing times.

3 Proposed Methodology

Effective machine learning can be hampered by a labor-intensive endeavor to reduce feature engineering outcomes in a representation of the data. It is feasible to find representative manners by learning from data rather than setting up acceptable features based on understanding by restoring low-level handcrafted features with learnt hierarchical characters and with the help of auto-encoders. Auto-encoders, unlike existing techniques, are utilized to recover classic sparse coding systems. Extracting feature representations from videos and learning a feature model is the same thing [4]. The use of numerous layers of hidden units in an auto-encoder can also be used to achieve hierarchical feature learning.

There are many methods for classifying normal and abnormal events in this work, but each method has its own advantages and disadvantages. The most accurate methods of existing technology are classifiers that are obtained by combining supervised and unsupervised classifiers and forming a fusion. Figure 2 depicts the architecture of CNN.

Max-fusion and SVM-fusion, Max-fusion is the method which compares all the predictions of classifiers and gives the highest probability for abnormality as result, whereas for SVM-fusion it is a method where it gathers the classifiers with the labels and rely on decision function for the result, and these results are measured for accuracy using ROC; ROC is receiver operator curve which compares the classification

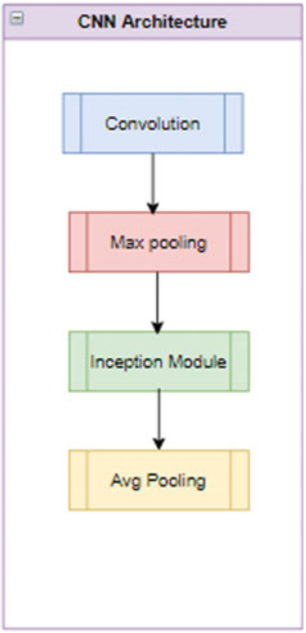


Fig. 2 CNN architecture

threshold value while plotting true positive rate and false positive rate, and another method is deriving the PRC; PRC is precision recall curves which gives more accurate value than ROC as it only gives the importance to the detected events, and it compares the detected events to the overall events, and overall performance will give by the AUC; AUC is area under the curve it is a common metric to measure the classifiers.

3.1 CNN Architecture

Another method is using CNN architecture in an approach which consists of convolutional neural networks, classifiers, dataset, and evaluation metrics. Convolutional neural networks will have different layers main three of them are convolutional, sub sampling, and fully connected layer through which we obtain the spatial features and temporal features to make the input for the classifiers and as for the classifiers, different classifiers are used such as SVM, Naive Bayes, KNN, K-clustering and result is obtained from all these algorithms. Max pooling will give the accurate model will be decided on the result of AUC which is area under curve and then dataset will be divided into training dataset and testing dataset normally 80 and 20 ratio will be used where necessarily both the dataset will contain normal and abnormal events. Average pooling will be done after the results; the sensitivity will be measured based on total predicted positive to the total positive events and also the negative events to the total negative events. An inception module is an image model block that aims to approximate an optimal local spare structure.

Spatiotemporal Recurrent Neural Network

Reconstruction of Input video Sequence:

Reframing: The size of video frames to fit into the model, resizing the each and every frame into common size in which all will be of same size.

Deconvolution: Process of filtering a signal to recreate it as original.

Temporal Decoder: Many temporal fragment inputs are given, though the required features are only extracted.

Temporal Encoder: The minimum number of features extracted and processed to classify into as many as possible clusters.

Figure 3 shows the work flow diagram unusual event detection such as traffic violations, robberies, and natural causes.

Convolution: Process of converting the value of one pixel to the weighted average of all pixels in the neighborhood.

RNN: Recurrent neural networks are a type of artificial neural network in which the connections between nodes create a directed graph over time, and it is the first algorithm to recall the input in its internal memory. Figure 4 depicts the spatiotemporal RNN architecture.

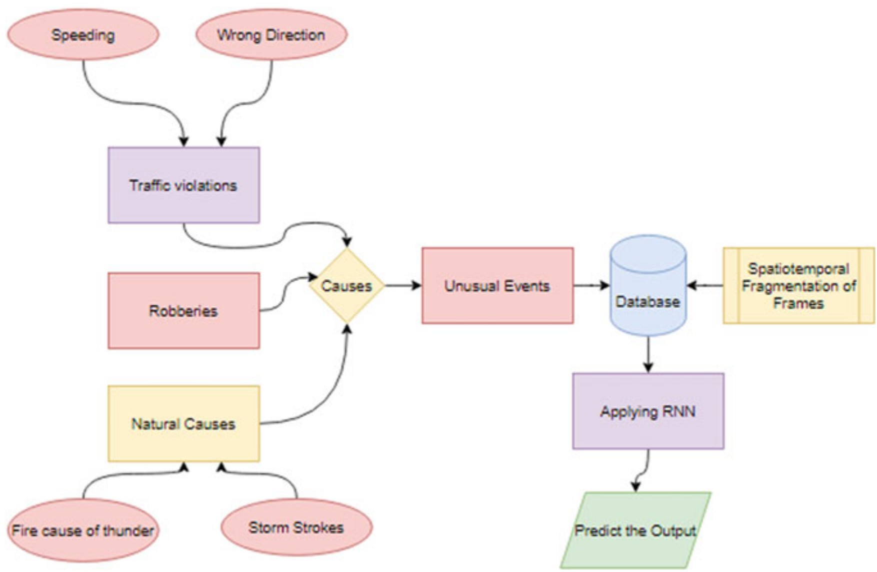


Fig. 3 Work flow diagram

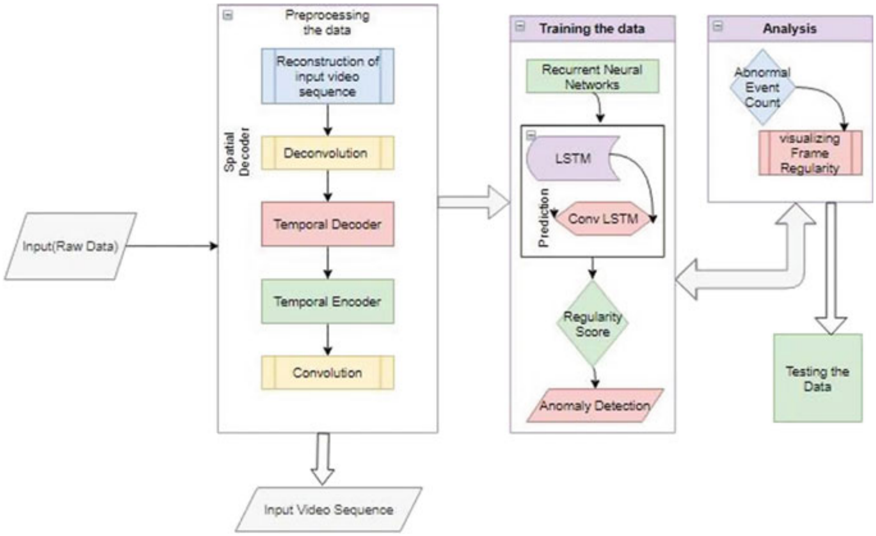


Fig. 4 Spatiotemporal RNN architecture

LSTM: Long short-term memory is a type of RNN; LSTM is capable of learning order dependence in the sequence prediction problems. LSTM gives the most control ability than RNN.

Regularity Score: Computing the reconstruction of error of a pixel's value of intensity at location (x, y) in video frame t, where the model is learned by LSTM convolutional auto-encoder.

Anomaly Detection: Anomaly detection is a technique to identify unusual patterns and behaviors, is also known as outlier analysis.

Abnormal Event Count: Abnormal events predicted to the actual abnormal events occurred, to find the accuracy.

Frame Regularity: Representing the frames which are more related the abnormal event occurred.

A data flow chart shows the way information flows through a process or system. It includes data inputs and outputs, data stores, and therefore the various sub-processes the info moves through. DFDs are constructed using consistent signs and representations to define many objects and their associations. Data flow diagrams visually represent systems and processes that might be hard to explain during a chunk of text. These diagrams can be used to map an existing system and make it better or to plan out a replacement system for implementation. Visualizing each element makes it easy to spot inefficiencies and produce the simplest possible system.

Figure 5 depicts the process of unusual event detection. First the datasets are collected and the images or frames are resized from the video by using spatiotemporal fragmentation. By applying recurrent neural network algorithm, the frames are

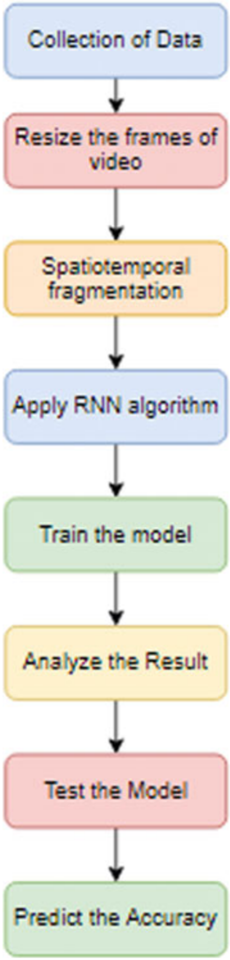


Fig. 5 Process of unusual event detection

trained and analyzed the result by testing the model. Finally, the unusual event was detected with higher accuracy.

The data obtained is raw data which is in the forms of complete video which is unprocessed and from different resources which probably do not have Meta data, video input will be converted into the form which trained model be able to accept for that, the raw data need to be aligned into proper fitting frames. To do so the video needs to be resized to 227 $\times$ 227 for each frame, thus every frame will be of the same size and also make sure that every input image is on the same scale values of the pixels need to be scaled in the range of 0 and 1 and also find the difference of each frame from the normalization of global mean image. To find the mean image calculate the average value of pixels at the place of each frame from the training dataset.

After obtaining the mean images and resizing the frames, the images must be converted to gray scale for dimensionality reduction, and normalizing the processed images to a zero mean and unit variance is required. The input of the trained model is processed video frames which are ten consecutive frames (video can be divided into frames based on the time limit give in seconds) which is known as strides basically give, here 1 s to get many numbers of strides, here the model parameters are huge, thus need the input which is training data to be huge, thus it is easier to generate the volumes, the frames with the strides make the sequence of the strides these can also be made in different methods like consecutive or odd or even, and take all three types to get more accuracy. In this way, the input of the model will be made to fit the model.

Algorithm 1: Spatiotemporal Recurrent Neural Network

```
Input: video_ frames from camera
Output: Usual or Unusual event
Monitoring_car_ movements_in_traffic_signal (video_frames)
pl ← Moving car in Red Signal _Detection()
    if (pl = = true)then
        send "alert"
        send "warning message to traffic police, relations"
        store "alert message in cloud"
    End if
q ← Robust move in Yellow Signal_Detection()
if(q = = true) then
        send "alert"
        store "alert message in cloud"
End if
driver_status_detection()
if((No licensed driver) = = True)then
        set "Risky"
else
```

Next step is to train the model with the common and regular patterns, to do that we have an architecture in which it contains mainly two parts they are spatial auto-encoder and temporal encoder and decoder, spatial encoder is for spatial structure of each and every video frame whereas temporal encoder and decoder is for temporal patterns of spatial structures. Auto-encoder has two phases they are encoding and decoding. First step is for dimensionality reduction by training the back-propagation algorithm from unsupervised learning through the reconstruction error minimization, and the result will be decoded from original input, encoder will extract many required features rather than common transformation methods.

Spatial convolution is about extracting the features using convolutional networks; convolution will store the spatial relation between the pixels in input frames. In mathematical form, convolution operations are done by using dot products for filters and regions in input. Traditionally, we use feed forward neural networks as we presume that all frames are rely on each other whereas temporal dependencies of input frames will be extracted using these convolutional networks. Recurrent neural networks

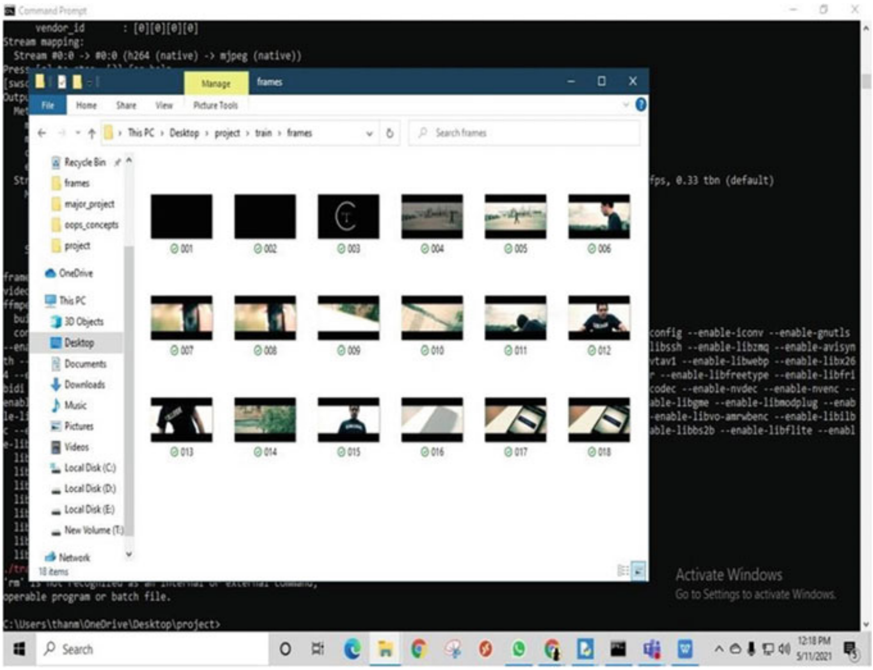


Fig. 6 Output of unusual event detection

and long short-term memory are used for mathematical derivations; however, final threshold value will be defined using Euclidean formulae, here we find the threshold values using mean square distance. Figure 6 depicts the output of unusual detection.

Algorithm 2: Red Signal Detection

```
Input: video_ frames from camera, trained_ dataset Output: Usual or Unusual event
f ←video_frames
For every frame f in video_frames
        For every frame p in trained_dataset
          if(f = = p)
          print "Unusual Event Detection"
          return "true"
    else
        print "Normal Event"

          End if
          End For
End For
```

4 Conclusion

The proposed system is very useful to take care of remote places where manpower cannot be inconvenient. Model needs to be tested with more real-time data with the installation to live data as till now it is trained only with the incidents already happened. Need to validate working while it keeps on running round the clock for minimum number of days. This will helpful to reduce the accidents or unpleasant incidents and makes it easy to investigate. In the future work, monitoring all the information using the proposed system and can access information to take further decisions with an immediate alert system to the nearby in charges and hospitals, police stations.

References

1. Wang H, Klaser A, Schmid C, Liu C (2011) Action recognition by dense trajectories. In: IEEE conference on computer vision pattern recognition (CVPR), pp 3169–3176
2. Xiang T, Gong S (2008) Incremental and adaptive abnormal behavior detection. Comput Vis Image Underst 111(1):59–73
3. Nguyen VQ, Anh TN, Yang HJ (2019) Real-time event detection using recurrent neural network in social sensors. Int J Distrib Sens Netw 15(6)
4. Loy CC (2019) Activity understanding and unusual event detection in surveillance videos. IEEE J 14(13):62–68
5. Loy C, Xiang T, Gong S (2010) Incremental activity modeling in multiple disjoint cameras. IEEE Trans Pattern Anal Mach Intell (TPAMI)
6. Chaquet JM, Carmona EJ, Fernández-Caballero A (2013) A survey of video datasets for human action and activity recognition. Comput Vis Image Underst 117(6):633–659
7. Singh R, Vishwakarma S, Agrawal A, Tiwari MD (2011) Unusualactivity detection for video surveillance. In: Proceedings of the first international conference on intelligent interactive technologies and multimedia. Association for Computing Machinery, New York, NY, USA, pp 297–305
8. Hao W, Zhang R, Li S, Li J, Li F, Zhao S, Zhang W (2020) Anomaly event detection in security surveillance using two- stream based model. Secur Commun Netw 2020(Article ID8876056)
9. Xue H, Liu Z (2013) Pedestrian classification based on improved support vector machines. In: International conference on intelligent networking and collaborative systems, pp 726–730

Improved Mental Health Monitoring Using HappyOSN in Online Social Networks

Poornima Nedunchezhian and Murugan Mahalingam

Abstract It is proven scientifically that the human mind behaves in accordance with the thoughts in subconscious mind. It is well known that the subconscious mind stores all the information that the human perceives. The HappyOSN is developed to improve the mental health of the online social network (OSN) person by monitoring the activity log of the persons. The HappyOSN firstly identifies the communication correlation. Secondly, the activity log is monitored, and the topics are text classified using PyTorch to identify the content category label. Thirdly, the subliminal repository is searched for the corresponding content category and is automatically posted in their OSN page. The subliminal content is the hidden frequency embedded in the audio or video content. Then, the emotion shift is monitored periodically. The Facebook, 2019 dataset is used. The accuracy, recall and F1-score metrics are considered. The recorded results showed significant improvement in all the considered metrics when compared with the results obtained from the existing feature ensemble model.

Keywords OSN—Online Social Networks · Subliminal · Frequency · Emotion shifts

1 Introduction

It is a well known that the human lives are created by themselves by the options they pick as decisions. The act of making decisions is guided by the human brain. The options made by humans are based on the ideas and the thoughts they think

P. Nedunchezhian (✉) · M. Mahalingam
Department of Information Security, SCOPE, Vellore Institute of Technology, Vellore, Tamil Nadu 632014, India
e-mail: poornima.n@vit.ac.inom

M. Mahalingam
e-mail: vp@valliammai.co.in

Department of Electronics and Communication Engineering, SRM Valliammai Engineering College, Kattankulathur, Kancheepuram (Dist.), Tamil Nadu 603203, India

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_36

subconsciously and consciously. The average Internet usage after COVID-19 has increased to a peak, and it will not reduce in the near future. The things we see and think about have more impact on our daily lives. Humans can change their life by focusing on positive things and the things they want rather than concentrating on negative things. Since people send most of their time online, we use the OSN as a key source. People generally take medical check-up yearly to check their health status. Similarly, it is important to consider their mental health status regularly. This research concentrates on the mental well-being of the humans. We propose an algorithm in the idea to monitor the mental health of humans through OSN and is named as HappyOSN.

It is scientifically proven by the researchers, that the things that are repeatedly seen are manifested knowingly or unknowingly. Hence, the proposed HappyOSN regulates the content shared in the OSN to take care of the mental health. The work progress is as follows. The OSN communication links are noted and recorded using the matrix factorization technique. Then, frequency of communication among humans is calculated periodically and recoded to find the communication statistics. Then, each person is assigned under three labels based upon their interests. The person interest is identified based on the things they search and the content they watch. This is achieved using the image classifier. Then, we attempt to alter their emotions using HappyEngine. The HappyEngine sends them content with subliminal boosters. The subliminal is the sound waves below 1000 Hz (Hz) that are unrecognized consciously by the human ear but it is identified subconsciously by the human brain. The positive affirmations are embedded as subliminal messages to the content shared by the HappyEngine to the depressed person. The depressed person is monitored for their emotion shifts periodically. Thus, the mental health of the persons is monitored, and hence, the system is named as HappyOSN.

The manuscript is organized as follows: Sect. 2 explains the literature survey. Section 3 detailed the overview on the working methodology of the proposed HappyOSN algorithm. Section 4 elaborates the design of the proposed HappyOSN algorithm. Section 5 explains the performance evaluation. The note on dataset description, performance comparison and evaluation results is explained. Section 6 concludes the paper along with future work.

2 Related Works

Adewole et al. [1] addressed the dark of social networks based on the taxonomy of features based on n-gram feature extraction, textual features and URL features. It performs crowd sourcing in the graph-based model and analyses using the supervised and unsupervised machine learning tools. Yilin et al. [28] developed similar research work to identify the malicious activity in social network. Another research work that works based on the text classification are done by Agarwal and Kausal [2]. It includes the feature extraction for the morphological and semantic analysis. The data preprocessing is done and the feature selection using feature extraction language (FXL) using the string similarity and corpus-based similarity. The feature loops (Floops) are identified for text classification. A fine-grained social network recommender system was developed by Aivazoglou et al. [3] based on sentiment analysis using the text processing. Another such model was developed by Alom et al. [4] based on the deep learning model for Twitter spam detection. Brailovskaia et al. [5], Chiu et al. [7], Chu et al. [8] proposed the depression discovery of persons. The automated models were developed based on the open-source classifiers. Classification is done on the content topics and the posts.

The researchers Cai et al. [6] and Wang et al. [25] developed the topic-based sentiment analysis for friend recommendation. Deng et al. [10] also developed the recommender based on the trust propagation. Another emotion analysis research work based on the community influence was noted by Chung and Zeng [9]. Zhang et al. [30] also performed emotion analysis using the posted content. Gibson and Trnka [11] proposed the social media that establishes safety by identifying the reason for stress. It focuses on providing the safety to teen age social media users by exhibiting care. Ju et al. [12] developed the structure-based positive influence maximization in social networks model which comprises the propagation function that identifies the path propagation probability by calculating the propagation increment values. Kadhim [13] gave a detailed survey on all the existing automated text classifiers. Kokciyan and Yolum [14] proposed the PriGuard algorithm. The sematic approach found the privacy breach in social networks. Le et al. [15] proposed the two methods to mine the Top-k details from the database, namely uncertain frequent patterns (UFPs) and Top-k UFPs (TUFP) mining. From the survey, the idea of picking the Top-k relevant application information is arrived. Poornima and Murugan [19] developed the natural disaster resilience approach (NDRA). It is a real-time application for quickest disaster recovery.

Another real-time application named SybilSort was developed in 2021 by the same researchers worked on the decision monitoring online friend recommendation system. Both the works focused on preventing the malicious registration in social networks, simple calculations to identify the malicious activity using the activity log and also confirmed the prediction. Rajesh Kumar et al. [20] developed a research work to predict the suicidal idea using the machine learning technique. The improved ensemble random forest (RF) algorithm was used which enhanced the baseline algorithm with improved data pre-processing and feature extraction technique. Sergio et al. [21] developed the research strategy that concentrated on predicting the early risk detection (ERD) of depressed people in social media. This worked in collaboration with the mental health doctors. It helped the depressed persons to receive the counselling without barriers. The work lacked automation and is more prone to human errors and is also time-consuming. The research work executed by Taridzo [22] identify the patients in depression automatically using the improved text classifier based on monitoring the activity. The day-to-day activity is monitored using their location tracking and by analysing their communication.

Tran et al. [23] developed the method that recommends the contents to the person based on their characteristics. The character analysis utilized the activity log. Wagner et al. [24] analysed the post-traumatic stress disorder (PTSD) and complicated grief (CG) of the family members of the suicidal persons in online social networks. It was executed based on the questionnaire-based survey. Xu et al. [26] modelled the online scholar friend recommendation system. The recommender gave the scholars the in-depth clarity about the topic in trend without manual searching. Yang et al. [27] developed VoteTrust algorithm which worked based on the reputation scores. The friend request is given along with the local repudiation score. The submitted score is compared with the global repudiation score to ensure integrity. Yin et al. [29] and Zhou and Chen [31] worked on identifying the most influential negative person in social networks using the correlation and matrix factorization.

3 Working of HappyOSN Algorithm

The OSN activity log is analysed to extract the connection statistics using the connection matrix. The adjacency matrix is formed using the modelled probabilistic graph model named Bayesian network. The developed model gives the matrix factorization of the OSN persons considering the joint probability. The Bayesian graph is modelled using the communication establishment to find the connection statistics. The evidence about the OSN activities is derived from the activity log. The negative association is identified and the more specific feature subsets are selected. The preprocessed data is then fed into the classifier for further analysis of the content view. The classifier helps to identify the content label. The subliminal repository is triggered and the emotion shift monitor tracks the emotion shifts. The overall working is in Fig. 1.

4 Design of Proposed HappyOSN Algorithm

4.1 Research Objectives

Objective 1: Preventing the depression of OSN persons, Objective 2: Monitor mental health passively, Objective 3: Track emotion shifts and Objective 4: Automate the mental health monitoring.

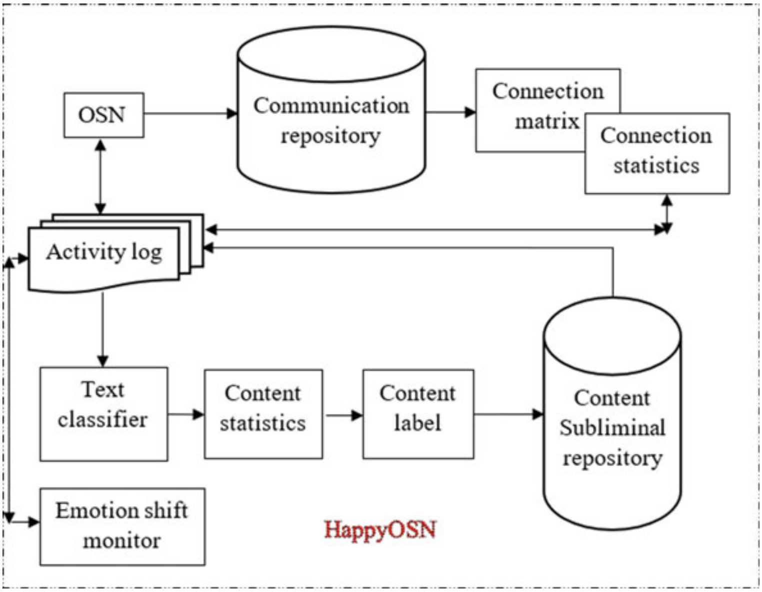


Fig. 1 Working of HappyOSN

4.2 System Model

The persons P in the OSN $\ddot{O}$ is represented as a collection of persons $P_i, P_j, \ldots, P_n$ where $k = 1, 2, \ldots, n$ and is represented in Eq. (1).

$$P = P_i, P_j, \ldots, P_n \tag{1}$$

The connection repository is given using the connection between the persons P_i and P_j in Ö using the matrix factorization. The connection matrix is written as shown in Eq. (2).

$$Þ_{P_iP_j} = \sum_{k=1}^{n} C_{P_iP_k} . C_{P_jP_k} \tag{2}$$

The matrix is plotted {True} if the communication exists between the persons else the connection matrix is plotted as {False} and is exhibited as Eq. (3). The $\overset{\vee}{\mathrm{C}}$orr correlation among the P_i and P_j is derived from the calculated ratio between the covariance $\overset{\vee}{\mathrm{C}}$ov and the variance var as shown in Eq. (4). The $\overset{\vee}{\mathrm{C}}\mathrm{ov}_{P_iP_j}$ is the mean of the product of the difference in similarity among P_i and P_j.

$$C_{P_iP_k} = \begin{cases} 1; \text{ if } C_{P_iP_k} = \text{True} \\ 0; \text{ if } C_{P_iP_k} = \text{False} \end{cases} \tag{3}$$

$$\overset{\vee}{\mathrm{C}}\,\mathrm{orr}_{P_i P_j} = \frac{\overset{\vee}{\mathrm{C}}\,\mathrm{ov}_{P_i P_j}}{\mathrm{var}_{P_i} * \mathrm{var}_{P_j}} \tag{4}$$

where the $\overset{\vee}{\mathrm{C}}\,\mathrm{ov}_{P_i P_j}$ is shown in Eq. (5) and the P_i and P_j content variance $C\mathrm{var}_{P_i}$ and $C\mathrm{var}_{P_j}$ is in Eqs. (6) and (7) and the elaborated $\overset{\vee}{\mathrm{C}}\,\mathrm{orr}_{P_i P_j}$ is in Eq. (8). The $\mathrm{var}_{\mathrm{P}_i}$ gives the association between the P_i and P_j. P_t refers the total number of persons in $\ddot{O}$.

$$\overset{\vee}{\mathrm{C}}\,\mathrm{ov}_{P_i P_j} = \left(\frac{\sum_{k=1}^{n} [C_{P_i P_k} - \mu(C_{P_{i,t} P_{k,t}})][C_{P_j P_k} - \mu(C_{P_{j,t} P_{k,t}})]}{\sqrt{\sum_{k=1}^{n} [C_{P_i P_k} - \mu(C_{P_{i,t} P_{k,t}})][C_{P_j P_k} - \mu(C_{P_{j,t} P_{k,t}})]}} \right)^2 \tag{5}$$

$$\mathrm{var}_{P_i} = \frac{[C_{P_i} - \mu(C_{P_t})]}{N} \tag{6}$$

$$\mathrm{var}_{P_j} = \frac{[C_{P_j} - \mu(C_{P_t})]}{N} \tag{7}$$

$$\overset{\vee}{\mathrm{C}}\,\mathrm{orr}_{P_i P_j} = \frac{\left(\frac{\sum_{k=1}^{n} [C_{P_i P_k} - \mu(C_{P_{i,t} P_{k,t}})][C_{P_j P_k} - \mu(C_{P_{j,t} P_{k,t}})]}{\sqrt{\sum_{k=1}^{n} [C_{P_i P_k} - \mu(C_{P_{i,t} P_{k,t}})][C_{P_j P_k} - \mu(C_{P_{j,t} P_{k,t}})]}} \right)^2}{\frac{[C_{P_i} - \mu(C_{P_t})]}{N} * \frac{[C_{P_j} - \mu(C_{P_t})]}{N}} \tag{8}$$

The consolidated connection correlation among 1000 persons is executed The PyTorch open-source text classifier is used for the proposed HappyOSN since it handles the dynamic architecture efficiently and the out of vocabulary words. The PyTorch is implemented by importing the PyTorch package which has PyText natural language processing (NLP) capacity for pre-processing the considered dataset. Table 2 gives the overall comparative analysis of the list of classifiers, namely support vector machine (SVM), Naive Bayes (NB), k-nearest neighbor, tensor flow (TF) and PyTorch classifiers. The classification time in seconds attributes for classification, accuracy, precision, recall and F1-score metrics. The FacebookGraphAPI dataset is used to extract the Facebook dataset using Python tokens. Then, the list of persons events is acquired using the graph.request() method from the activity log. Then, the Facebook dataset is downloaded to the destination location as the .csv file. The pre-processing of dataset is done by tokenizing, converting the uppercase to lowercase, then the batch size is allocated for each round.

Algorithm NegativeAssociationRuleMining
*NegativeAssociationRuleMining (Input: Training_Dataset:*ꝱŦ *, Positive_Association*: $PA(A \rightarrow B)$ *: Output: Negative_Association*: $NA{\sim}(A \rightarrow B)$*,* $PA(A \rightarrow B)$*,R)*
Start
    Set NA= ȹ
    Ƒ $= 0$
    *While (*ꝱŦ)
        If $Check\_frequency\_PA: A \rightarrow B$
            Set Ƒ $++$
            Update_PA_rules.Add(PA, F*)*
        Endif
    EndWhile
//Calculate Rank R
    Foreach a in PA

$$R = \begin{cases} 1; if\ F > (1/4)\Sigma(F) \\ 2; if\ F = (1/4)\Sigma(F) \\ 3; if\ F < (1/4)\Sigma(F) \end{cases}$$

        Assign (PA, R)
    EndForeach
//Derive NA from PA
    Foreach w in PA

$$Find\ \ NA = \{\frac{P(AUB)}{P(A)P(B)}\} - \{\sim(P(A).P(B)\}$$

        Update NA
    End Foreach
End

The collected dataset ꝱŦis pre-processed, and the negative association NA is identified. The feature selection algorithm is also depicted. The PA is identified and the frequency Ƒis noted and updated. Then, the PA repository is scanned, and the NA is removed if identified. The FS is extracted considering the NA and PA repository.

*Algorithm FeatureSelection (Input: Training_Dataset:*ꝱŦ *, Positive_Association*: $PA(A \rightarrow B)$ *, Negative_Association*: $NA{\sim}(A \rightarrow B)$*, R; Output: Featureset: FS)*
Start
    Set FS= ȹ
    *While(*ꝱŦ*≠ȹ)*
        Evaluate $PA = PA(A \rightarrow B)$
            Foreach asso in PA
                if $NA{\sim}(A \rightarrow B)$
                    Pass
                else
                    Update FS
                Endif
            End Foreach
    EndWhile
End

The sample of ten rounds of training and testing validation is done. The accuracy and the loss information are recorded. The test results show the better performance with optimized time, accuracy and loss. The classification loss mentioned in Table 3 is recorded based on Eq. (9).

$$\text{Loss} = \frac{-1}{\text{total}} \sum_{k=1}^{\text{total}} P(O)\ln(\mathrm{P}(C)).P(1-O)\ln(\mathrm{P}(1-C)) \tag{9}$$

The feature set is extracted for labelling Ļthe category of the person by classifying the topic of text, image and video content they watch using the activity log. The feature set is extracted using the monitored evidences. The probability that the P_i belongs to P_i is based upon the evidences E_z monitored from the activity log. This is equated as Eq. (10).

$$P(\mathrm{P}_i|\text{Ļ}) = \prod_{z=1}^{n} P(\mathrm{P}_i|E_z) = P(\mathrm{P}_i|E_1), P(\mathrm{P}_i|E_2), P(\mathrm{P}_i|E_3), \ldots, P(\mathrm{P}_i|E_n) \tag{10}$$

The probability that the feature term T_i that exists in $\text{Ļ}_j$in the document D_k is the feature term selection function in the overall D_k. $P(T_i|\text{Ļ}_j|D_k)$in Eq. (11) gives the probability of the frequency of the term.

$$P(T_i \mid \text{Ļ}_j \mid D_k) = \sum_{i,j,k=1}^{m,n,p} P(T_i, \text{Ļ}_j, D_k).\log \frac{P(T_i, \text{Ļ}_j, D_k)}{\sqrt{P(T_i).P(\text{Ļ}_j).P(D_k)}} \tag{11}$$

The $\frac{P(T_i,\text{Ļ}_j,D_k)}{\sqrt{P(T_i).P(\text{Ļ}_j).P(D_k)}}$is the information gain which gives the mutual relevance between the T_i, $\text{Ļ}_j$and D_k. Then, the relevancy score $RS(T_i|\text{Ļ}_j)$of the person emotion to the $\text{Ļ}_j$is calculated to find the content statistics as represented in Eq. (12). The emotion is classified as $\{\mathrm{Sad}_{P_i}||\mathrm{Happy}_{P_i}\}$.

$$\mathrm{RS}(T_i|\text{Ļ}_j) = log\sqrt{\frac{P(T_i,\text{Ļ}_j)}{\sqrt{P(T_i).P(\text{Ļ}_j)}}} \tag{12}$$

$$\underset{i,j=1}{lim}^{n,m}\text{Ļ}_{j_{P_i}} = \underset{i,j=1}{lim}^{n,m}\{Sad_{P_i}|Happy_{P_i}|Fuzzy_{P_i}\} \tag{13}$$

$$\begin{pmatrix} \mathrm{Sad}\,\delta_{P_i} = \frac{\sum_{a=1}^{N} \mathrm{prob}(\mathrm{RS} \geq \mathrm{SA}(P_i(a)))}{N} \\ \mathrm{Happy}\,\delta_{P_i} = \frac{\sum_{a=1}^{N} \mathrm{prob}(\mathrm{RS} \geq \mathrm{HA}(P_i(a)))}{N} \\ \mathrm{Fuzzy}\,\delta_{P_i} = \frac{\sum_{a=1}^{N} \mathrm{prob}(\mathrm{RS} \geq \mathrm{FA}(P_i(a)))}{N} \end{pmatrix} \tag{14}$$

If the P_i is labelled as Sad_{P_i}, then the content subliminal repository is invoked. The PyPI is used by installing pip in Python to store the collection of subliminal audio and video. The person $RS(T_i|\text{Ļ}_j)$is calculated, then the final decided label is then decided, then the relevant subliminal audio and video are posted to the person as recommended information. Then, the emotion is analysed to find the emotion shifts.

$$Var_{Ⴑ_{jP_i}} = \frac{(Ⴑ_{jP_i} - \overline{Ⴑ_j})^2}{Total - 1} = \begin{cases} 0; & if < constant \\ 1; & if \geq constant \end{cases} \tag{15}$$

Equation (15) gives the variance of the prediction results. The $Ⴑ_{jP_i}$ and the mean of the predicted $\overline{Ⴑ_{jP_i}}$ is derived to calculate the variance.

$$cd = \int_{i,j=1}^{n} \frac{1}{Var_{Ⴑ_{jP_i}} \sqrt{2\Pi}} e^{-1/2} \left(\frac{\sqrt{\frac{1-n}{n^2}} . \left(\overline{Var_{Ⴑ_{jP_i}}}\right) - \left(A(Var_{Ⴑ_{jP_i}})\right)}{n-1} \right) . dx \tag{16}$$

The cd between the predicted label to the $Var_{Ⴑ_{jP_i}}$ to the actual $Ⴑ_{jP_i}$;$A(Var_{Ⴑ_{jP_i}})$is given in Eq. (16). The existence of the deviation D_N degrades the overall results, and hence, it needs to be minimized or to be limited in order to optimize prediction $f(\text{opt})$. This is expressed as Eq. (18). The equation is derived from the Frobenius norm by minimizing the squared errors to half.

$$D\_N = \begin{cases} 1; & \text{if cd} > k \\ 0; & \text{if cd} < k \\ F; & \text{if cd} == k \end{cases} \tag{17}$$

$$f(opt) = \frac{1}{2} \sum_{i=1;j=1}^{N;M} ||Var_{Ⴑ_{jP_i}} - \overline{Var_{Ⴑ_{jP_i}}}||^2 + \frac{1}{2} \sum_{i=1;j=1}^{N;M} ||Ⴑ_{jP_i} - \overline{Ⴑ_j}||^2 - D\_N \tag{18}$$

Table 1 gives the statistics of the prediction optimization. The group of persons are analysed for identifying the $Var_{Ⴑ_{jP_i}}$ and the calculated cd. The derived D_N is

Table 1 Prediction optimization statistics

Statistics				
Persons	$Var_{Ⴑ_{jP_i}}$	cd	D_N	$f(\text{opt})$
10,000	0.9213	0.2753	0	0.2261
20,000	0.9649	0.9444	1	0.7486
30,000	0.7255	0.8501	1	0.6569
40,000	0.9144	0.5124	F	0.8162
50,000	0.7301	0.5496	1	0.4419
60,000	0.6334	0.0873	0	0.0854
70,000	0.7792	0.9154	1	0.2470
80,000	0.9603	0.1786	0	0.1104
90,000	0.6523	0.2631	0	0.0748
100,000	0.3964	0.8564	1	0.1864

plotted as {0||1||F} based on Eq. (17). The calculated and the minimized variance is plotted as f(opt). Equations (19) and (20) are given to find the send subliminal audio send subliminal video for P_i on $Ḽ_j$.. Then, after the period t, the emotion is analysed and labelled again as $\left\{\text{Sad}_{P_i} \middle| \text{Happy}_{P_i} \middle| \text{Fuzzy}_{P_i}\right\}$.

$$SSA(P_i(Ḽ_j)) = \begin{vmatrix} SSA(P_i(Sad_{P_i})) \\ SSA(P_i(Happy_{P_i})) \\ SSA(P_i(Fuzzy_{P_i})) \end{vmatrix} \tag{19}$$

$$SSV(P_i(Ḽ_j)) = \begin{vmatrix} SSV(P_i(Sad_{P_i})) \\ SSV(P_i(Happy_{P_i})) \\ SSV(P_i(Fuzzy_{P_i})) \end{vmatrix} \tag{20}$$

5 Performance Evaluation

5.1 Dataset Description

The total of 1,00,000 Facebook 2019 persons are considered with the total communication link of 5,28,123 with no artificially synthesized label. Table 2 explains the details of the PyTorch training and testing. The training and testing are executed on the topics extracted from the mostly viewed image, text, audio and video on the labels $\left\{\text{Sad}_{P_i} \middle| \text{Happy}_{P_i} \middle| \text{Fuzzy}_{P_i}\right\}$. The word size of 50 is considered for a batch while training and testing.

Table 2 PyTorch Classification

Average accuracy percentage = 99.369						
Label	Sad$_{P_i}$	Happy$_{P_i}$	Fuzzy$_{P_i}$	Sad$_{P_i}$	Happy$_{P_i}$	Fuzzy$_{P_i}$
Average audio topics label of rounds				*Video topics label of rounds*		
Sad$_{P_i}$	98.11	0	0	99.54	0	0
Happy$_{P_i}$	0	96.14	2.00	0	97.62	5.0
Fuzzy$_{P_i}$	1.89	3.86	98.00	0.46	2.38	95.2
Total	100	100	100	100	100	100
Text topics label of rounds				*Image topics label of rounds*		
Sad$_{P_i}$	99.98	0	0	96.31	0	0
Happy$_{P_i}$	0	96.98	4.10	0	100.00	0.67
Fuzzy$_{P_i}$	0.02	3.02	95.90	3.69	0.00	99.33
Total	100	100	100	100	100	100

Table 3 Classification accuracy

	HappyOSN			Feature ensemble model		
	Precision	Recall	F1	Precision	Recall	F1
Sad_{P_i}	0.97	0.98	0.98	0.88	0.86	0.89
Happy_{P_i}	0.95	0.92	0.99	0.85	0.79	0.82
Fuzzy_{P_i}	0.94	0.97	0.97	0.77	0.85	0.79
Total	0.95	0.96	0.98	0.83	0.83	0.83

5.2 Evaluation Results

Table 2 gives the average classification accuracy percentage of 99.369 on the audio, video, text and image topics classified using the PyTorch. The average results on each round are aggregated to find the average classification accuracy. Table 3 gives the overall precision, recall and F1-score values of the proposed HappyOSN over the already existing feature ensemble model.

5.3 Performance Comparison

Table 4 showcases the overall error err noted over each round during different time periods $\overline{\mathrm{T}}_x, \overline{\mathrm{T}}_y$ and $\overline{\mathrm{T}}_z$. From the tabulation, it is concluded that the proposed HappyOSN has least error compared to the feature ensemble model. Figure 2 gives the comparison of precision, recall and F1-score of the proposed HappyOSN to the existing feature ensemble method.

6 Conclusion and Future Work

The subliminal messages are the scientifically proven method to alter once thought process. The application of subliminal information is used by the proposed HappyOSN to monitor and track the emotion shifts of the OSN person. The statistics says that the person spends around 145 min per day in OSN. The category of the content they view manifests their emotions and lives. In order to maintain the good mental health, the HappyOSN has the subliminal message repository. The person is categorized as sad, happy and fuzzy based on the topic of content they watch. The person with sad or fuzzy nature is motivated passively using the hidden subliminal messages by embedding them inside the content they watch. Almost all the persons concentrate only on the physical health and have no or very least concentration on the mental health. The novelty of the proposed HappyOSN is it saves human lives and also improves the mental health. Since the motivation is done passively, the overall

Table 4 Error extraction

Round	HappyOSN			Feature ensemble model		
	err ($\overline{\mathrm{T}}_x$)	err ($\overline{\mathrm{T}}_y$)	err ($\overline{\mathrm{T}}_z$)	err ($\overline{\mathrm{T}}_x$)	err ($\overline{\mathrm{T}}_y$)	err ($\overline{\mathrm{T}}_z$)
1	1.11e−10	4.37e−10	3.21e−11	2.41e−5	3.47e−5	6.44e−5
2	4.32e−11	4.63e−11	2.07e−14	5.91e−4	5.11e−4	4.76e−4
3	3.83e−12	5.82e−11	7.29e−12	2.12e−4	5.14e−4	5.65e−4
4	6.29e−12	1.22e−11	2.82e−11	5.23e−5	6.43e−6	8.36e−5
5	8.45e−11	1.67e−12	7.63e−10	7.44e−5	2.49e−5	9.35e−6
6	4.98e−12	3.32e−12	8.92e−16	4.11e−4	1.77e−4	3.22e−4
7	5.21e−11	3.43e−11	3.38e−13	5.81e−5	5.76e−5	2.46e−5
8	1.87e−12	2.80e−12	7.28e−13	9.81e−4	4.22e−4	2.92e−6
9	2.52e−14	8.49e−12	6.91e−17	8.12e−4	2.73e−4	1.02e−4
10	4.65e−15	1.27e−17	3.61e−16	1.23e−5	7.66e−6	5.32e−5

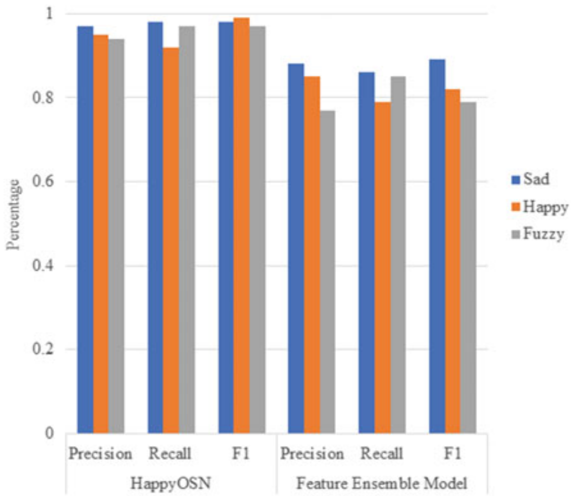


Fig. 2 Comparison of HappyOSN and feature ensemble model

design will be a disaster if the malicious person handles the subliminal messages and will be extended as the future work.

References

1. Adewole K, Anuar N, Kamsin A, Varathan K, Razak S (2017) Malicious accounts: dark of the social networks. J Netw Comput Appl 79:41–67
2. Agrawal H, Kaushal R (2016) Analysis of text mining techniques over public pages of Facebook. In: IEEE 6th international conference on advanced computing (IACC), Bhimavaram, pp 9–14
3. Aivazoglou M, Roussos AO, Margaris D, Vassilakis C, Ioannidis S, Polakis J, Spiliotopoulos D (2019) A fine-grained social network recommender system. Soc Netw Anal Min 10(1)
4. Alom Z, Carminati B, Ferrari E (2020) A deep learning model for Twitter spam detection. Online Soc Netw Media 18:100079
5. Brailovskaia J, Forkmann T, Glaesmer H, Paashaus L, Rath D, Schonfelder A, Juckel G, Teismann T (2019) Positive mental health moderates the association between suicide ideation and suicide attempts. J Affect Disord 245:246–249
6. Cai C, Xu H (2019) A topic sentiment based method for friend recommendation in online social networks via matrix factorization. J Vis Commun Image Representation 65:102657
7. Chiu C, Lane H, Koh J, Chen A (2020) Multimodal depression detection on instagram considering time interval of posts. J Intell Inf Syst
8. Chu SH, Lenglet C, Schreiner MW, Klimes-Dougan B, Cullen K, Parhi KK (2018) Anatomical biomarkers for adolescent major depressive disorder from diffusion weighted imaging using SVM classifier. In: Annual international conference IEEE engineering in medicine and biology society
9. Chung W, Zeng D (2020) Dissecting emotion and user influence in social media communities: an interaction modeling approach. Inf Manage 57(1):103108
10. Deng S, Huang L, Xu G, Wu X, Wu Z (2017) On deep learning for trust-aware recommendations in social networks. IEEE Trans Neural Netw Learn Syst 28(5):1164–1177

11. Gibson K, Trnka S (2020) Young persons's priorities for support on social media: "It takes trust to talk about these issues." Comput Hum Behav 102:238–247
12. Ju W, Chen L, Li B, Liu W, Sheng J, Wang Y (2020) A new algorithm for positive influence maximization in signed networks. Inf Sci 512:1571–1591
13. Kadhim A (2019) Survey on supervised machine learning techniques for automatic text classification. Artif Intell Rev 52(1):273–292
14. Kokciyan N, Yolum P (2016) PriGuard: a semantic approach to detect privacy violations in online social networks. IEEE Trans Knowl Data Eng 28(10):2724–2737
15. Le T, Vo B, Huynh VN (2020) Mining top-K frequent patterns from uncertain databases. Appl Intell 50:1487–1497
16. Notredame CE, Grandgenevre P, Pauwels N, Morgieve M, Wathelet M, Vaiva G, Seguin M (2018) Leveraging the web and social media to promote access to care among suicidal individuals. Front Psychol 9
17. Notredame C, Morgiève M, Morel F, Berrouiguet S, Aze J, Vaiva G (2019) Distress, suicidality, and affective disorders at the time of social networks. Curr Psychiatry Rep 21(10)
18. Poornima N, Murugan M (2020) Natural disaster resilience approach (NDRA) to online social networks. J Ambient Intell Humanized Comput 12(5):5651–5678
19. Poornima N, Murugan M (2021) SybilSort algorithm—a friend request decision tracking recommender system in online social networks. Appl Intell. https://doi.org/10.1007/s10489-021-02578-x
20. Rajesh Kumar E, Rama Rao KV, Nayak SR, Chandra R (2020) Suicidal ideation prediction in twitter data using machine learning techniques. J Interdisc Math 23(1):117–125
21. Burdisso SG, Errecalde M, Montes-y-Gomez M (2019) A text classification framework for simple and effective early depression detection over social media streams. Expert Syst Appl 33:182–197
22. Taridzo C (2014) Text classification to automatically identify online patients vulnerable to depression. In: Pervasive computing paradigms for mental health. Lecture notes of the institute for computer sciences, social informatics and telecommunications engineering, vol 100, pp 125–130
23. Tran V, Hwang D, Nguyen N (2018) Hashtag recommendation approach based on content and user characteristics. Cybern Syst 49(5–6):368–383
24. Wagner B, Keller V, Knaevelsrud C, Maercker A (2015) Social acknowledgement as a predictor of post-traumatic stress and complicated grief after witnessing assisted suicide. Int J Soc Psychiatry 58(4):381–385
25. Wang R, Li Z, Cao J, Chen T, Wang L (2019) Convolutional recurrent neural networks for text classification. In: International joint conference on neural networks (IJCNN), pp 1–6
26. Xu Y, Zhou D, Ma J (2019) Scholar-friend recommendation in online academic communities: an approach based on heterogeneous network. Decis Support Syst 119:1–13
27. Yang Z, Xue J, Yang X, Wang X, Dai Y (2016) VoteTrust: leveraging friend invitation graph to defend against social network Sybils. IEEE Trans Dependable Secure Comput 13(4):488–501
28. Yilin W, Jiliang T, Jundong L, Baoxin L, Yali W, Clayton M, Neil O, Yi C (2017) Understanding and discovering deliberate self-harm content in social media. In: International world wide web conference committee (IW3C2), pp 93–102
29. Yin C, Zhang X, Liu L (2020) Reposting negative information on microblogs: do personality traits matter? Inf Process Manage 57(1):102106
30. Zhang X, Li W, Ying H, Li F, Tang S, Lu S (2020) Emotion detection in online social networks: a multilabel learning approach. IEEE Internet Things J 7(9):8133–8143
31. Zhou Q, Chen G (2020) An efficient victim prediction for Sybil detection in online social network. IEEE Access 8:123228–123237

Dr. Poornima Nedunchezhian is currently working as an Assistant Professor at Vellore Institute of Technology, Vellore. She is a, Life Member of ISTE and CSI. She has more than 25 publications to her credit in reputed International/National journals and conferences. In particular, her recent interests include Online Social Networks Security, Emotion Dynamics, Recommendation Systems and Data Analytics

Murugan Mahalingam graduated in E&CE from the University of Madras in 1989. Received his Masters in E & TC Engineering (Microwave) in 2001 and Ph.D. in 2010 respectively from the University of Pune. Presently he is serving as the Professor of ECE Dept. and the Principal of SRM Valliammai Engineering College, Kattankulathur, Tamil Nadu. Having about 32 years of teaching, his fields of interests are Antennas, Microwave, SatCom, OptiComm and EMI&C. He is a Fellow of IETE, Life Member of ISTE, IEI, ISOI, SEMCEISSI and Member of CSI and ISCA. Over 1250 papers are published in the International/National repute Conferences/Journals and written four text books. He received the prestigious ISTE-Anna University National Award for an Outstanding Academic (2016) and IETE-Prof. K. Sreenivasan Memorial Award (2020).

Detection of Brain Tumor Using Neuro-Fuzzy Classifier

G. Tamilmani and S. Sivakumari

Abstract In the biomedical image processing, the total variation of image processing is the initial step to be executed. Selecting proper filter is a vital task. The filter output can be a variation of image without degradation at the edges with ROI. The next step of variation is segmentation. The primary objective of segmentation is to extract the region of interest from image for particular utilization. Due to its difficult structure, the pathological tissue segmentation from typical brain MRI is a hard with consuming more time. Many approaches considered so far are time-consuming and contain certain limitations. To deal with this, a segmentation method has to be detected. The classification mode is a noteworthy task of tumor identification. To classify the brain tumor, neuro-fuzzy classifier is used.

Keywords Neuro-fuzzy classifier · Magnetic resonance imaging · Region of interest

1 Introduction

Fuzzy system is the rule-based expert system utilizing fuzzy rules including fuzzy inference. Fuzzy rule signifies commonsense of direct mode, knowledge with skills, where knowledge is subjective, ambiguous, vague, or **contradictory. The** proposed model evaluation metrics demonstrate its proficiency of diagnosing brain tumor MR images likened to sensitivity, specification, and accuracy. Figure 1 represents neuro-fuzzy systems. Here, the fuzzy system has three major components: (1) fuzzy input–output variables (2) computed its fuzzy values, and (3) set of fuzzy rules fuzzy inference approach. The input data is crisp; similarly, output value is anticipated to crisp also, and then the estimation of fuzzification rule, defuzzification inference mode has been used over the fuzzy rules of IF mode denotes, denotes, and 'y' denotes 'B.'

G. Tamilmani (✉) · S. Sivakumari
Avinashilingam Institute of Home Science and Higher Education for Women, Coimbatore, India
e-mail: taman.maya@gmail.com

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_37

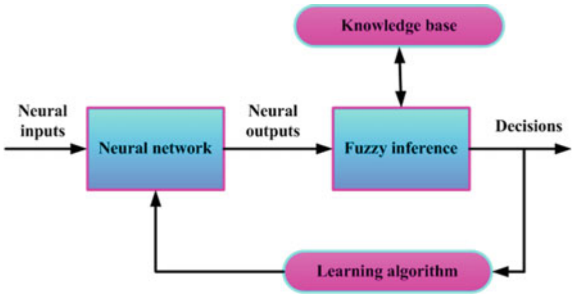


Fig. 1 Neuro-fuzzy system

Fuzzy logic method is

- Determine the linguistic variables with terms (initialization)
- Set the MF (initialization)
- Set the rule base (initialization)
- Modify crisp input data as fuzzy values in terms of MF (fuzzification)
- Estimate the rules on rule base (inference)
- Integrate every rule outcome (inference)
- Modify the output data as the values of non-fuzzy (defuzzification).

The fuzzification is defined as detecting procedure the membership degrees μ () and μ () input data ‘ ’ and ‘ ’ belong to the fuzzy sets and at previous portion of the fuzzy rule. The amounts of input data compatibility with the conditional components of a rule are calculated by fuzzification. After the process of fuzzification, the rule evaluation is deemed. It manages single values of membership degrees μ () and μ () and also generates output MF ‘B.’ Defuzzification is defined as computing procedure. Here, the predicted result for this variable is the numeric value of the single output for the fuzzy output variable based on the membership function. By processing data samples, the neuro-fuzzy system is fuzzy systems that determine the properties based on neural networks. The neuro-fuzzy combines the features of NNs with fuzzy logic on complementary mode to prevent its drawbacks. NNs are strong in representing the relationships amid the input and output variables. The system enhancement is based on the reorganization capabilities of neural network, which provides specified accuracy for certain applications.

1.1 Classification Utilizing Neuro-Fuzzy Classifier

NFC is nothing but NEFCLASS, because it is small for neuro-fuzzy classification. It is employed to data analysis through neuro-fuzzy schemes. By supervised learning, it could learn fuzzy rules with fuzzy sets. The features of NEFCLASS are:

- It represents the scheme of fuzzy classification.
- It incrementally learns the rules of fuzzy classification.

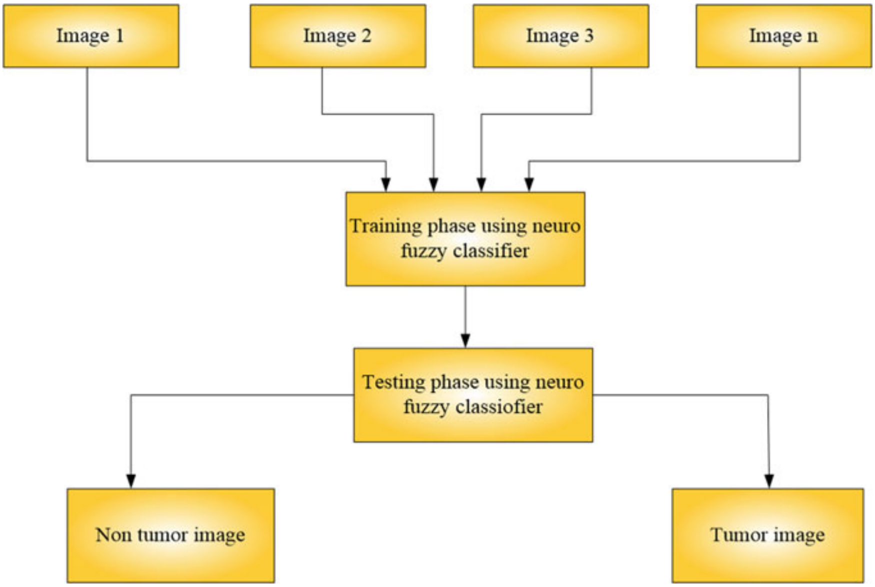


Fig. 2 Classification using NFC

- It terms fuzzy sets depending on simple heuristics.
- The learning approach does not affect the basic fuzzy classifier semantics.

This system is explained with the help of fuzzy classification rules; also, the learning capacity of neural networks is provided by hybrid neuro-fuzzy method for fuzzy inference systems [1]. The input and output data are deployed through the learning approach for training process, and the membership functions of surgeon-type fuzzy inference scheme are estimated to validate the efficiency. Figure 2 shows the classification using NFC.

2 Module Description

2.1 *Fuzzy Neural Networks (FNN) with Neuro-Fuzzy Systems (NFS)*

Fuzzy neural networks (FNNs) are a specific type of neural networks made up of fuzzy neurons. Fuzzy neural networks have a distinct characteristic of exhibiting synergic collaboration between fuzzy concepts and neural networks. FNNs are classified based on the connection of neurons in the network. The general type of FNNs is feed forward, in which the neurons are grouped in layer-by-layer fashion, and an input signal travels the entire network in one-way direction. In the same manner, fuzzy neurons have not specified connection; also, its networks are referred to as nets without a feedback connection. Also, there exists a specified type of neural networks called recurrent networks, in which the signal passes the entire network at two directions. This type of networks possesses a dynamic type of memory and thus represents its state in dynamic manner. A neuro-fuzzy network (NFN) is basically trained with the help of a learning algorithm provided by an intelligent framework. The integration of neural network with fuzzy logic thus provides robust, effective, and simply to understand for several applications.

Depending upon the complex of specified problem, the structure and number of layers in the architecture of hybrid systems can vary. A simple hybrid network can have one or two hidden layers in their structure. The hidden layer is responsible for fuzzy inference system, fuzzification, generating rules, and defuzzification. Here, the first hidden layer performs the fuzzification process with organization of data for training, selection of optimal number of samples and identification of bias values, whereas the last layer performs the process of defuzzification and thereby provides the final outputs or decisions of the model.

2.2 Neuro-Fuzzy Systems

NFSs have attracted many investigators in several scientific and engineering fields because of its proficient learning mechanisms and reasoning efficiency. In the field of artificial intelligence, the term neuro-fuzzy refers to the integration of artificial neural networks with fuzzy logic. The basic architecture of neuro-fuzzy systems is illustrated in Fig. 3 as follows.

In fuzzy systems, the estimation of a set of membership functions and rules is normally a time-consuming task. In this concern, several learning algorithms are incorporated into the fuzzy systems. This makes the membership function estimation process easier. The neural networks with efficient learning algorithms support better tuning of fuzzy systems. The NFS combines the learning power of ANN, thereby explicitly the knowledge representation of fuzzy inference systems. The NFS is used in majority of applications that include data analysis, data classification, detection of imperfections, and decision-making process.

2.3 Fuzzy Rule-Based Classification System (FRBCS)

FRBCS is used to solve the classification issues. In the image processing systems, the various kinds of problems are generated in the image processing system that is clarified by using this classifier. In this, the class is attained as with a reassigned class set is given as for object. The object is the portion of some feature space, and then the classifier is used to maintain the appropriate class [2]. Then this classifier deals with the fuzzy rules. The types of the fuzzy rules are formed on the basis of

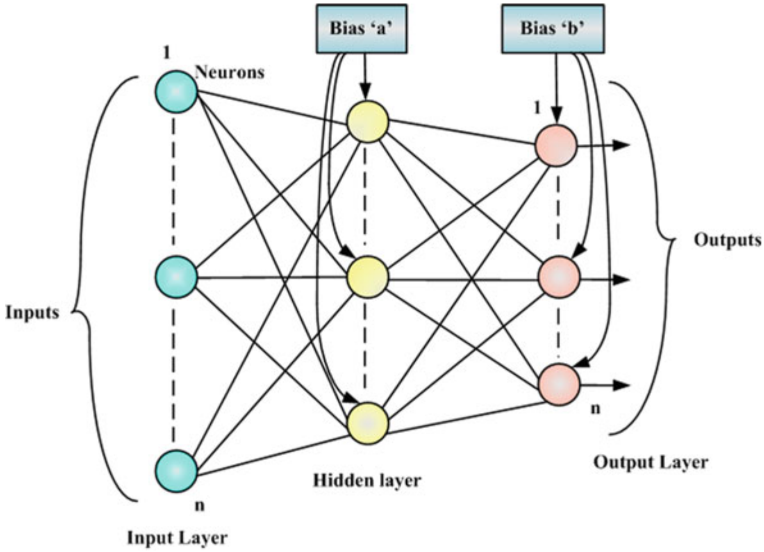


Fig. 3 Typical architecture of NFSs

the NN, decision tree, and FDT. The classifier generates the set of the fuzzy rules then the process is called fuzzy rule-based classification mode. The fuzzy variables are defined by fuzzy rules antecedent, which presents the computational feasibility. Training models with classifier of this property can solve the issue. Then this model gives the set of the new samples. To solve this problem, a Fuzzy ID3 algorithm is used in the classifier. This fuzzy approach can build the tree based on the learning process. This algorithm is used to get the less uncertainty and to get the more information's by using the fuzzy entropy (Fig. 4).

Then each path tree consists of the three rules, and also, the leaf node consists of the rule weight (RW) for the number of the sets. In this, WR_i denotes the jth weight rule and handle the fuzzy assurance value as AS_i and this value is equal to the WR_i. Then to perform the classification task, fuzzy rule-based method is used. In this, the fuzzy rules based on the set of the class and the certainty degree with the consequent are given as

$$T_l : if\ a_1\ is\ X_1^l \text{ and} \ldots \text{and} \ldots a_M\ is\ X_M^l \text{ then } B \text{ is } S_i \text{ with } t_l \tag{1}$$

where t_l is denoted as the certainty degree on classification with S_i class and it belongs to fuzzy substance and constraints with fuzzy antecedent.

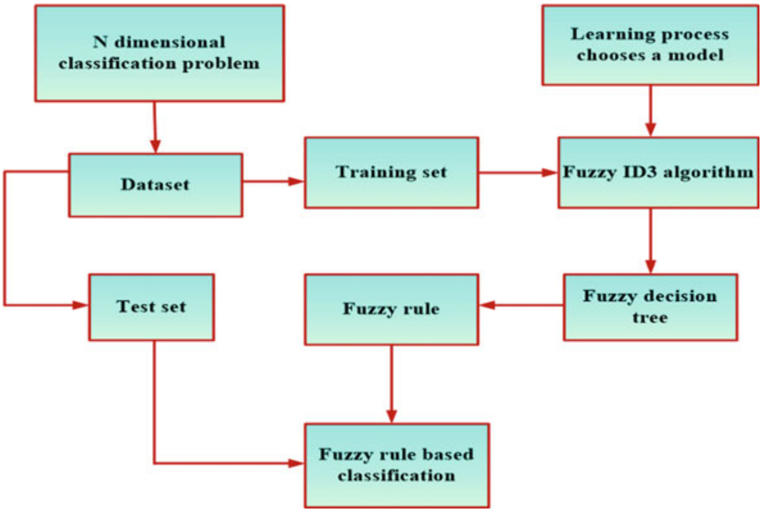


Fig. 4 Block diagram for fuzzy algorithm to solve the classification problem using combined FID3 and FRBC

2.4 Fuzzy Interactive Dichotomizes 3

The adaptation of the decision tree is called fuzzy decision tree (FDT) by the function of the fuzzy logic. There are various kinds of the adapted from the fuzzy logic and to build the FDT. The tree is created; also, decision rules have been reached on the basis of every path from the root through the leaves in the trees. Fuzzy interactive dichotomizes 3 (Fuzzy ID3) is mostly used in the classification tree to build the algorithm. This algorithm separates the data set on the basis of data quality that is chosen by means of a measure named as information gain in terms of fuzzy entropy.

From this, let us assume the training set that has the o samples with $A_o = (A_{o1}, \ldots, A_{om})$ and contain the oth sample of the training set. Here, A_{oj} represents the value of jth attribute and it is given as with the oth training samples. Then the every samples are belong to the class, and it is given as $b_o \in S = \{(S_1, S_2, \ldots, S_m)\}$, where n is represented as the number of the class and the computational complexity is solved. In this, M is represented as the fuzzified patterns and the m attributes are given as $\{(X_1, X_2, \ldots, X_m)\}$. In this, l is assumed as $(1 \leq l \leq m)$. Then the attributes X_l and the n_l values of the fuzzy subset are given as $\{(X_{l1}, X_{l2}, \ldots, X_{lml})\}$. In this, S is represented as the classification with the target attributes and the n values are given as $S_1, S_2, \ldots, S_m$. Then the symbol $N(.)$ has been employed for representing the cardinality of particular fuzzy set and sum of membership values with fuzzy subset. Then the step of fuzzy ID3 induction process is mentioned below:

Step 1. To generate a root node

In this, root node generates and that root node consists of every data set. Then every data is fuzzified as well as every membership degree of the values is equal to 1 through the data to initialize.

Step 2: Selecting the attributes

In this, the attributes are selected by five steps; they are:

Step 2a: Compute the relative frequencies

This is to compute the relative frequencies through the respective classes such as $S_i(i = 1, 2, \ldots, m)$ for every linguistic label, and the label is given as $X_{lj}(j = 1, 2, \ldots, m)$.

Then the relative frequency equation is given as:

$$o_{lj}(i) = \frac{N(X_{lj} \cap S_i)}{N(X_{lj})} \tag{2}$$

Step 2b: Compute its fuzzy classification entropy

Compute its fuzzy classification entropy through the linguistic labeled as $X_{lj}(1, 2, , \ldots, n_l)$. Then the fuzzy entropy equation is given as:

$$\text{Entrophy}_{lj} = \sum_{i=1}^{n} o_{lj}(i) \log(o_{lj}(i)) \tag{3}$$

Step 2c: Compute the average fuzzy classification entropy

Compute the average fuzzy classification entropy EN_l for the every attribute.

$$EN_l = \sum_{i=1}^{n_l} \frac{N(X_{lj})}{N(X_{lj})} \text{Entrophy}_{lj} \tag{4}$$

Step 2d: Select the attribute

The attributes () are selected by the process of increasing the profit information P_l . Then the attribute equation is given as

$$\text{attributes} = \max_{1 \le l \le m} (P_l), \ \text{where } P_l = \text{EN}_l = \text{entrophy}_{lj} \tag{5}$$

Step 2e: Assign the selected attribute

Assign the selected attribute by the functions of the root node including linguistic labels denoted by the tree's branches (Fig. 5).

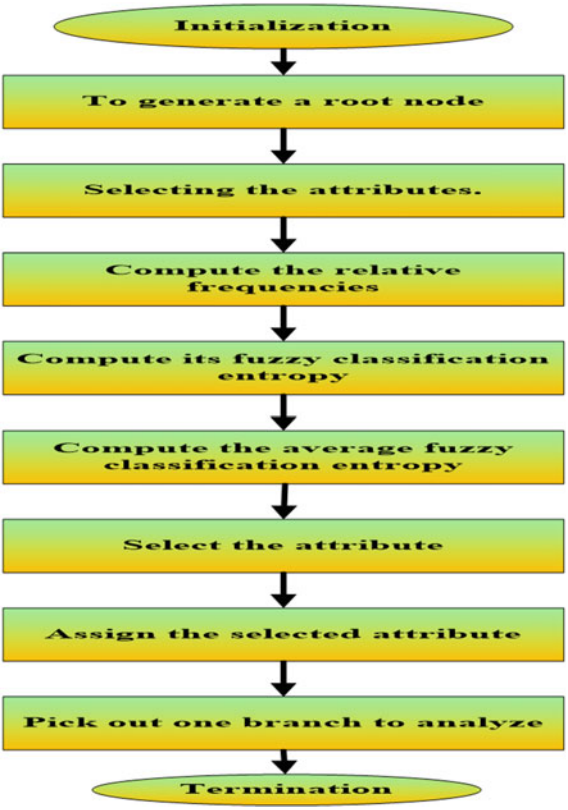


Fig. 5 Flowchart for fuzzy interactive dichotomizes 3 (Fuzzy ID3)

Step 3: Pick out one branch for analyzing

To analyze the branch by picking one of the branches from the above step. While analyzing the branches, it will remove the branches containing no information's and then estimates the relative frequencies from Eq. (4). While detecting the frequencies is in above the threshold value φ_t, then all the attributes are extended in this branch and then terminate the branch as the leaf. Otherwise, choose the attribute from the above steps that are not extended this branch with least average fuzzy classification entropy is selected from Eq. (5) as a new decision node to the branch, then add its linguistic labels as candidate's branches for analysis. On every leaf, each class contains its relative frequency.

Step 4: Termination

Repeat step 3 to analyze the branches. If it is lack of candidate branches, then the decision tree is summed by the process of the rules based on the FIDM. While finishing the above steps, a rule is generated and the rules T_r are given below,

Rule T_r:

Let consider the FID3 rules are given as: T_l : $if\ a_1\ is\ X_1^l$ and ... and ... $a_M\ is\ X_M^l$ then B is S_i with t_l. Here $a = (a_1, a_2, \ldots, a_m)$ with n-dimensional pattern vector. This vector function denotes fuzzy set as X_{lj}, and class label is represented as S_{lj} and WR_i are denoted as the rule weight. In this fuzzy algorithm, every leaf node consists of the rule weights. To every class, the rule weights originated from the relative frequency. This is mentioned in step 3.

3 Result and Discussion

3.1 Efficiency

To estimate the efficiency of the proposed neuro-fuzzy classifier model at brain tumor, the confusion matrix for segmented image is produced, and comparison between the neuro-fuzzy classifier model outputs along its relative original image is carried out based on these created confusion matrices. By these produced confusion matrices, the performance measures like accuracy, sensitivity, precision, F-score, positive predictive values, negative predictive values, specificity, false reject rate, false accept rate, for determining how exactly the brain tumor is categorized. There are four statistical indices are computed from the produced confusion matrix and also utilized for evaluating the proficiency of the proposed model: True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN).

- True Positive (TP): Benign properly recognized into benign.
- True Negative (TN): Malignant properly recognized into malignant.
- False Positive (FP): Malignant imperfectly recognized into benign.
- False Negative (FN): Benign imperfectly recognized into malignant.

3.1.1 Accuracy

The correct classification accuracy or rate is the classification efficiency suitable to total count of classification tests. This classification mode of brain cancer is implemented at several normal and abnormal MRI images, and Eq. 6 is utilized to compute accuracy.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100\% \tag{6}$$

3.1.2 Sensitivity

Sensitivity calculates the proportion of actual positives that are properly recognized. Calculating the sensitivity uses Eq. 7,

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \tag{7}$$

3.1.3 Specificity

Specificity is also named as TN rate. It calculates the proportion of actual negatives, which is properly recognized. Calculating the specificity uses Eq. 8,

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100\% \tag{8}$$

3.1.4 F-score

F-score calculates the accuracy of model in a data set. It is utilized to estimate binary classification systems that classify examples as 'positive' or 'negative.' To calculate, the F-score uses Eq. 9,

$$\text{F - score}\frac{\text{TP}}{\text{TP} + \frac{1}{2(\text{FP}+\text{FN})}} \times 100\% \tag{9}$$

3.1.5 Positive Predictive Values (PPV)

Positive and negative predictive values imply proportions of positive and negative outcomes at statistic and diagnostic tests, which implies TP and TN outcomes, respectively. The ideal PPV value using ideal test represents 1 (100%), and worst possible value is zero.

The positive predictive values are obtained using Eq. 10,

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\% \tag{10}$$

3.1.6 Negative Predictive Values (NPVs)

The negative predictive values from time to time are known as negative predictive agreement. An NPV ideal value using ideal test is 1 (100%), and poor feasible value is zero.

The NPVs are estimated from Eq. 11,

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}} \times 100\% \tag{11}$$

3.1.7 False Reject Rate (FRR)

FRR is calculated as a fraction of positive scores falling below the threshold. The value of threshold in the range of 0 and 1. FRR is computed by Eq. 12

$$\mathrm{FRR} = \frac{\mathrm{FN}}{\mathrm{TP} + \mathrm{FN}} \tag{12}$$

3.1.8 False Accept Rate (FAR)

FAR is calculated as a fraction of negative scores exceeding the threshold.

$$\mathrm{FAR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{13}$$

3.2 Result

By testing various methods on the gray images, we finally concluded that the neuro-fuzzy classifier provides the accurate results comparing with the other data mining or classification methods like support vector machine and artificial neural network. Here the percentage of classification accuracy is more compared to the other methods which give the better accuracy. The values of the results are listed below which shows very clearly that the proposed method is more efficient when comparing with the other methods (Table 1).

Table 1 Experimental results

Evaluation parameter	Classification using artificial neural networks	Classification using support vector machine	Classification using neuro-fuzzy classifier (proposed)
Specificity (%)	85.27	86.15	88.05
Sensitivity (%)	86.11	88.73	94.03
Accuracy (%)	93.05	95	98
F-Score (%)	81.04	82.89	86.3
PPV (%)	76.54	77.78	79.74
NPV (%)	91.67	93.33	96.72
FRR (%)	13.89	11.27	5.97
FAR (%)	14.72	13.85	11.94

4 Conclusion

Classification is the major important step of brain tumor identification in the MR images. The biggest challenge is to ensure the classification of the tumor type with the time frame and also measure the experimental decision. Here, efficient brain MRI image classification technique using neuro-fuzzy networks is proposed. The features are extracted by the penalized fuzzy c means approach as well as the optimized feature is selected by GA along the joint entropy. The performance of the classifiers like neuro-fuzzy classifier and CNN classifiers with various kernels are estimated in GRBF kernel; it activates efficient likened to the other kernels. The accuracy of classification is found in the CNN classifier with GRBF kernel.

References

1. Lefohn AE, Cates JE, Whitaker RT (2003) Interactive, GPU-based level sets for 3D segmentation. In: Conference on medical image computing and computer assisted intervention (MICCAI), pp 564–572
2. Chan FHY, Lam FK, Poon PWF, Zhu H, Chan KH (1996) Object boundary location by region and contour deformation. IEEE Proc Visual Image Process 143(6):353–360
3. Saad A, Smith B, Hamarneh G, Möller T (2007) Simultaneous segmentation, kinetic parameter estimation, and uncertainty visualization of dynamic PET images. In: Springer on medical image computing and computer assisted intervention (MICCAI), vol 4792/2007, pp 726–733
4. Amini L, Soltanian-Zadeh H, Lucas C (2003) Automated segmentation of brain structure from MRI. Proc Intl Soc Mag Reson Med 11
5. Barr A, Lenoble M, Carugno G, Centro S, Charpak G, Cruickshank G, Lewiner J (2001) Imaging brain tumors using a multi-wire gamma camera and Thallium-201. In: 2001 IEEE nuclear science symposium conference record, vol 1, pp 452–456
6. Yazdan-Shahmorad A, Soltanian-Zadeh H, Zoroofi RA (2004) MRSI brain tumor characterization using wavelet and wavelet packets feature spaces and artificial neural networks. In: The 26th annual international conference of the IEEE engineering in medicine and biology society. IEMBS '04, vol 1, pp 1810–1813

7. Mortamet B, Zeng D, Gerig G, Prastawa M, Bullitt E (2005) Effects of healthy aging measured by intracranial compartment volumes using designed MR brain database. In: International conference on medical image computing and computer-assisted intervention, pp 383–391
8. Bricq S, Collet C, Armspach JP (2008) Unifying framework for multimodal brain MRI segmentation based on Hidden Markov Chains. Med Image Anal 12(6):639–652

A Systematic Study About EEG Signal Data and Computer Aided Models for the Diagnosis of Alzheimer's Disease

P. Saroja and N. J. Nalini

Abstract Dementia is brain disorder that can show impact on patient memory loss and impaired mental ability, such as speaking and thinking. Based on the several reasons, the dementia is occurred that considers Alzheimer's disease, vascular dementia, Dementia with Lewy bodies, etc. Early medication may help dementia patients to overcome the memory loss and thinking skills. Various techniques are also used to manage this disease with behavioral issues. Sometimes, the medications also cannot change the brain that can cause dementia. 65% to 85% dementia patients have chances to convert Alzheimer's disease. Electroencephalography (EEG) is the test that can calculates the electric activities that occur in brain. This test may help experts to estimate the status of the disease. EEG is also helps to diagnose the brain disorders such as dysfunction of brain (encephalopathy), dysfunction stroke, and sleep disorders. In this paper, the performance of various algorithms is analyzed by applying benchmark datasets and also real-time datasets for diagnosing and detection of Alzheimer's disease using EEG signals are reviewed. This helps researchers to select appropriate techniques according to their need.

Keywords Alzheimer's · Dementia · Electroencephalography (EEG)

1 Introduction

In 1906, the German physician Dr. Alois Alzheimer discovered Alzheimer's disease (AD) for the first time. AD is one of the challenges for healthcare in the twenty-first century. From the past, many people across the world suffering with AD. Generally, 70% of people greater than 65 years are living with AD. This is one of the chronic and irreparable that affects the brain cells and causes damage to the regular functioning

P. Saroja (✉) · N. J. Nalini
Department of Computer Science and Engineering, Annamalai University, Chidambaram, India
e-mail: pathapati.saroja@gmail.com

N. J. Nalini
e-mail: njncse78@gmail.com

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_38

of the brain. There is no proper treatment for this disease. Detection of AD in the early stages may reduce the panic situation of the patient, and sometimes, this can prevent the increase of this disease in the patient brain. Magnetic resonance imaging (MRI) and positron emission tomography (PET) are the two major tests that can detect and diagnose AD in the early stages based on the abnormal condition in the brain. Experts can identify the condition of the patient and the stage of the disease. Machine learning (ML) algorithms play a significant role in detecting and diagnosing AD with the integration of huge dimensional multimodal neuroimaging data.

From all over the world, approximately 3% of men and women between the ages of 65 to 75 are living their lives with AD [1, 2]. Every year 3,65,000 new cases of AD are diagnosed. There is no strong reason to cause this disease. By 2050, the number of persons suffering from AD may increase up to 153 million across the world. In this paper, the author focused on detecting and diagnosing AD in the early stages with ML algorithms. Several combinations of feature extraction and preprocessing techniques are used to diagnose this disease. The ML algorithms are applied to EEG signal datasets to analyze the stage of the disease.

To diagnose the AD in the early stages, some significant factors impact on patient.

- The alert signs of the AD in the early stages are memory loss, difficult in doing daily and regular tasks, issues in talking, i.e., language, problems with abnormal thinking, very poor or reduced judgment, misusing of things, sudden change in behavior or mood, drastic change in personality, loss of creativity. The symptoms of Alzheimer's disease are confusion, disturbances in short-term memory, problems with attention and spatial orientation, personality changes, language difficulties, unexplained mood swings.

The organization of the paper is as follows. Section 2 describes diagnostic tests on Alzheimer's disease. Section 3 explains the study of EEG signals of the human brain. Section 4 discusses the different types of preprocessing methods on EEG data. Section 5 discusses the different features extraction techniques. Section 6 describes various machine learning algorithms to detect Alzheimer's disease (AD). Section 7 describes experimental results, and the conclusion is included in Sect. 8.

2 Diagnostic Tests on Alzheimer's Disease

Detecting and diagnosing the Alzheimer's disease (AD) is based on the tests conducted to the patients. Based on health conditions of the patient and tests, the AD can be identified by the experts. Electroencephalogram (EEG) is one of the significant tests that can take patient sample analyze the status of the patient. Other brain imaging tests such as CT scan, MRI, PET, SPECT, CSF examination are also used. In the early stages, it is very difficult to diagnose the AD. An unpredictable diagnosis is naturally done when cognitive impairment compromises daily doing activities, if the person is still living very independently. That person can recover from MCI issues,

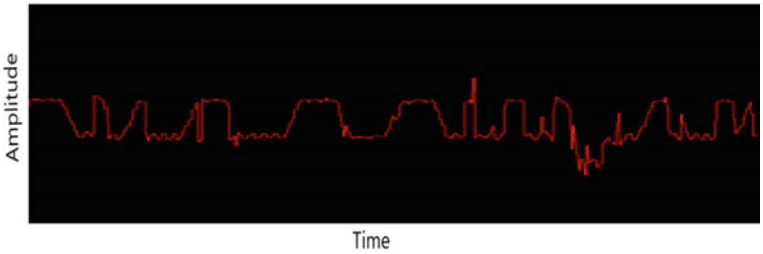


Fig. 1 Sample EEG signal obtained when the patient is in a laughing state

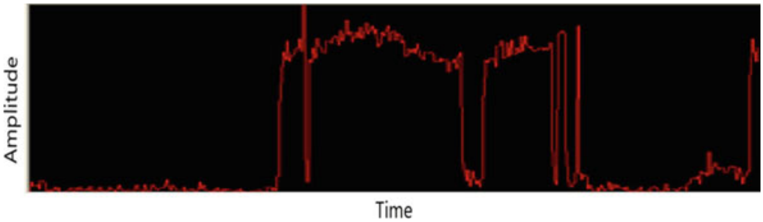


Fig. 2 EEG signal diagnosing Alzheimer's disease (AD)

such as loss of memory by improving the cognitive and non-cognitive variation, preventing the possibility of separate living (Figs. 1 and 2).

3 Study of EEG Signal

Electroencephalogram (EEG) is the method that can record the activities of brain. From many years, EEG is used to diagnose and detect the Alzheimer's disease. In human brain, activities are analyzed by polymorphic slow waves which are represented as theta or delta range, frontal intermittent rhythmic delta activity (FIRDA), and other EEG outputs in other patients. In present days, several new techniques are developed by using machine learning (ML) and deep learning (DL) to analyze the brain activities and status of brain by using EEG signal.

4 Preprocessing Methods on EEG Data

Electroencephalography (EEG) signals are used to provide the sensitive information to study the function of the brain and its disorders. These signals are recorded to contaminate by several noises that makes the analysis of signals very tedious. It is very important to remove the noises from the EEG input signals. This problem may obtain in the collection of patients. In brain, several factors show the impact based on movements of muscles, sounds of heart, eyes blinking, antique ocular, the additional noise is obtained from the movements of eyeballs which is very difficult. Several methods and techniques are developed and proposed to detect and diagnose the EEG signals.

Jeena Joy et al. [9] show the comparison between various DL algorithms to denoise the Alzheimer's disease (AD). In this domain, the diagnosing process prevents the noise by brink in the wavelet domain. DWT provides the multiple analyses with efficient solutions.

Janett et al. [10] proposed a unique model for denoising the complex EEG signals. The artifacts are removed by using the B-Spline by using the independent component analysis (ICA) and experiments show that performance of B-Spline Mutual Information Independent Component Analysis (BMICA) is better. Prasad et al. [11] proposed the new thresholding filter by using to denoising the EEG signal by utilizing the wavelet packets.

5 Feature Extraction Techniques

Various techniques such as independent component analysis (ICA), fast Fourier transform (FFT), and wavelet transformations (WT) are the feature extraction methods are used to extract the accurate information from the EEG signal [20]. Separation of blind source is one of the better examples of ICA. This can be developed by using the ICA which is used to find the signals to divide the noise from brain signals. From the computation point of view, ICA is very efficient method that can measure the high-quality EEG signals [21].

The following steps to implement the ICA with EEG signals.

- Subtracting the mean by using Center '*x*' by '*s*.'
- Whiten '*x*.'
- '*w*' represents the random initial value for the demixing matrix.
- '*w*' is calculated as new value.
- Initialize '*w*.'
- If the algorithm is not converged, then it returns to step 4.
- '*w*' and '*x*' to get the independent source signals.

$$\boldsymbol{S} = \boldsymbol{W}\boldsymbol{x} \tag{1}$$

To decompose the covariance matrix the eigenvalue is used and this is used to whitening the input signal. Equation 2 represents the mathematical formula:

$$\tilde{\boldsymbol{x}} = \boldsymbol{E}\boldsymbol{D}^{-1/2}\boldsymbol{E}^{\boldsymbol{T}}\boldsymbol{x} \tag{2}$$

'D' represents the diagonal matrix of eigenvalues, and 'E' represents the orthogonal matrix of eigenvectors.

Wavelets transforms are majorly divided into two classes: Continuous wavelet transform (CWT) and discrete wavelet transform (DWT). In CWT, huge energy is generated continuously based on the frequency bands. By using the WT, the non-stationary signals are initialized by the time–frequency domain. With the integration of variable window size, various advantages are obtained such as energy is extracted, the most significant disadvantage is cluster that selects the actual method to reduce the noise in application [22].

The frequency bands are represented in Eq. 3.

$$\gamma(s, \tau) = \int f(t)\psi s, \tau * (t)d \tag{3}$$

where * represents the tedious conjugation.

From equation, $f(t)$ is decay into a set of regular functions Ψs, $\tau(t)$, which are called as wavelets. The term and τ, represents the scale and translation

$$f(t) = \int \int \gamma(s, \tau)\psi s, \tau(t)\mathrm{d}\tau\mathrm{d}s \tag{4}$$

$$\psi s, \tau(t) = \frac{1}{\sqrt{s\psi\left[\frac{t-\tau}{s}\right]}} \tag{5}$$

's'—scaling factor,

'τ'—translation factor and

The factor $\sqrt{s}$ initializes the normalization for energy among the several scales.

DWT: For representation of the signal, DWT is one of the techniques that did not modify the data in a signal. The series of wavelets is just an inspected form of the constant wavelet change and the information given by it is profoundly excess to the extent the generation of the signal is distressed.

Fast Fourier transform (FFT): This technique mainly applied on EEG data, and this is also mathematical method. Power spectral density (PSD) is the prediction method that measures the characteristics of the extracted EEG signals that are analyzed for detecting the Alzheimer's disease (AD).

Mathematical formula of a Fourier series is:

$$x(t) = \frac{a_0}{2} + \sum_{k=1}^{\infty}(a_n \cos(\omega kt) + b_n \sin((\omega kt)) \tag{6}$$

The signal $x(t)$ is integral part and the interval [0, T] and is an angular ω periodic with period T, t is a time variable, frequency and a_0, a_n, b_n are Fourier coefficients.

The Fourier coefficient can be obtained by formulas (Table 1):

Table 1 Preprocessing techniques and algorithms for detection of Alzheimer's disease

Author name and year	Preprocessing techniques	Algorithms	Performance measures
G. Fiscon et al. (2018) [12]	Time–frequency transforms	Tree-based classifier C4.5	Accuracy
A. T. Khan et al. (2018) [13]	Asymmetric least squares and Savitzky–Golay transformation	Partial least square (PLS) and principal component analysis (PCA)	Accuracy
S. Siuly et al. (2020) [14]	Piecewise aggregate approximation (PAA) for dimensionality reduction	Extreme learning machine (ELM)	Accuracy and execution time
S. Nobukawa et al. (2020) [15]	Functional connectivity, estimated by phase lag index (PLI)	Support vector machine (SVM)	Accuracy
M. Ismail et al. (2019) [16]	FFT transformed to extract brainwave subbands (theta, alpha, and beta)	Event-related potential (ERP)	Accuracy
H. A. Gonzalez et al. (2019) [17]	Independent component analysis (ICA)	CNN classifier	Accuracy
Silvia Marino et al. (2019) [18]	1-hidden layer MLP	Support vector machine (SVM)	Accuracy
Guilherme Tavares et al. (2019) [19]	Proposed feature selection (FS)	Ensemble learning algorithm	Accuracy, F1-score, and features

$$a_n = \frac{2}{T} \int_0^T x(t)\cos(\omega k t)\mathrm{d}t, \quad k = 0, 1, 2, 3, \ldots, N$$

$$b_n = \frac{2}{T} \int_0^T x(t)\sin(\omega k t)\mathrm{d}t, \quad k = 0, 1, 2, 3, \ldots, N$$

6 Machine Learning (ML) and Deep Learning (DL) Methods

ML and DL are most widely used algorithms to detect and diagnose the dementia by using EEG signals. This algorithm plays the significant role in processing the complex EEG signals. Researchers use this ML and DL algorithms to detect various diseases.

6.1 Supervised Learning Algorithms

With the supervised algorithms, the EEG signals are processed based on the predicted values. This will classify the classification and regression.

- Classification can be used for problems where the output variables are categorized. Support vector machine, decision trees, nearest neighbor, discriminant analysis, neural network, etc. come under classification.
- Regression is used for the problems where the output values are continuous or real values such as salary, weight, and dollars. Examples of regression are linear regression, logistic regression, etc.

6.2 Unsupervised Algorithms

Unsupervised algorithms are used when the information used to train is neither classified nor labeled. It is categorized into clustering and association.

- Clustering divides the several items based on the similarity function. The main aim of this is to find the structure or pattern that collects the undivided data. Several clustering algorithms are also present, and these algorithms are applied on EEG signals dataset to detect the dementia and AD.
- Association is a rule-based algorithm; it helps to discover interesting patterns among the variables in large databases. Examples are Apriori, FP growth, etc.

6.3 Deep Learning Techniques

- ***Convolutional Neural Networks (CNN)***: CNN is one of the deep learning (DL) algorithms that will take a sample (any patient sample in any format such as image, EEG signals) as input and gives the accurate output according to the input. CNN consists of three layers such as a convolutional layer, pooling layer, and fully connected layer. In the first stage, the input EEG signal is identified, and training is given to the algorithm. The output layer of the hidden weights appears in several forms. The output of this algorithm consists of several numerical data such as binary format numbers, for example, patient sample classification, normal = 0, abnormal = 1, then the algorithm shows the classification output.
- ***Recurrent Neural Networks (RNN)***: RNN is most widely used in many deep learning domains to process the complex issues. RNN is mainly based on intra- and inter-layer connections for time series classification. It is very tedious task to train RNN. In RNNs, the back-propagation is modified to function that consists of iteration in the graphs. This approach extracts the features based on the sequential EEG signals to recognize the AD. By using forward connections, this will create the network that allows back-propagation to process as normal, and the cost is

also limiting the recurrent connections. Thus, this algorithm can process complex datasets such as EEG signals to detect the AD and dementia.

- ***Long Short-Term Memory networks (LSTM)***: LSTM is one of the deep learning approaches that consist of fully connected layers and activation layers, which are built in to encode the relations among the features and detect the next stage of Alzheimer's disease. In this algorithm, the data is transferred by using the next layer which is passed by using activation function. Anyway, the recurrent connection is present in the layer which is not belongs to activation function. In LSTM, lack of activation function is one the disadvantages, and also, this will prevent the vanishing gradient issue. Thus, this algorithm can be used to detect the AD by using EEG signals as input.
- ***Stochastic Gradient Descent (SGD)***: This model shows the difference in error function which is calculated by using EEG sample in dataset. This is most widely used and very easy to implement for training EEG datasets that fit into the memory. This method used to measure the weights very accurately; this is generally needs a huge amount of calculation time because the selected EEG signal data needs to change at every update. These methods solve the speed and accuracy problems. This algorithm SGD extracts the features randomly from the overall data for fast and most frequent updates. Several updated SGD algorithms are also extended to improve the performance of detecting and diagnosing the AD.

7 Experimental Results

The experiments are conducted by using Python programming language. Python has very powerful packages such as numpy, keras, pandas, and sklearn which are most important packages that can used in development of integrated learning algorithm to detect and diagnose the AD in patients. These algorithms are applied on various bench mark datasets. The performance measures, namely sensitivity, specificity, F1-score, and accuracy, the performance of the system are estimated (Fig. 3).

- **TP**: The predicted sample is positive, and it is true.
- **TN**: The predicted sample is negative, and it is true.
- **FP**: The predicted sample is positive, and it is false.

True Positive (TP)	**False Positive (FP)**	**False Positive Rate (FPR)**
False Negative (FN)	**True Negative (TN)**	**False Negative Rate (FNR)**

Fig. 3 Performance measures

- **FN**: The predicted sample is negative, and it is false.
- **Precision**: From the samples, the overall positives predicted.

$$\text{Precision (P)} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- **Recall**: From all the positives, the percentage of positives predicted.

$$\text{Recall (R)} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- **Specificity**: From the overall samples, the correct negative predictions are divided by total number of negatives. The values are in between 1(best) to 0(worst).
- **F1-Score**: This is the weighted average of precision and recall.

$$\text{F1} - \text{Score} = \frac{2 \times \text{P} \ \times \text{R}}{\text{P} \ \times \text{R}}$$

- **Accuracy**: It is the measure that predicts the all accurate predictions that are divided by overall number of the dataset (Table 2, Fig. 3 and Table 3).

8 Conclusion

This paper mainly focused on detecting and diagnosing of the Alzheimer disease in early stages by using EEG signal samples. Various ML and DL algorithms are applied on several datasets to diagnose the AD disease. Various feature extraction techniques and preprocessing techniques are discussed in this paper. Many authors select the suitable processing method for signal processing according to the requirements. Classification of EEG signals in DL is in early stages. To improve the performance of DL algorithms, the feature extraction and preprocessing algorithms are integrated to increase the accuracy. This research is carried out by using EEG signal datasets with the ensemble DL algorithms which shows the high performance in future. Example of a Computer Program from Jensen K., Wirth N. (1991) Pascal user manual and report. Springer, New York.

Table 2 Various feature extraction techniques and algorithms used in Alzheimer's disease

Author name and year	Feature extraction method	Algorithm	Performance measures (%)
Albaqami et al. (2020) [23]	Wavelet packet decomposition (WPD)	Gradient boosting, decision trees	Accuracy (86.60), sensitivity (81.8), specificity (90.8)
Ge Q et al. (2019) [24]	Maximal overlap discrete wavelet transforms (MODWT)	Discrimination framework	Accuracy (93.15), AUC (98.01), F-measure (94.20)
Pholpat Durongbhan et al. (2018) [25]	Fast Fourier transform (FFT)	K-nearest neighbor (KNN)	Sensitivity (73.90), specificity (86.90), and accuracy (84.10)
G. Filcon et al. (2018) [26]	Discrete Fourier and wavelet transforms (DFWT)	Decision tree classifier (DTC)	Accuracy (91.8), precision (91.77), sensitivity (91.89), specificity (91.15), F-measure (91.17)
Deepa et al. (2017) [27]	Fast Fourier transform (FFT), discrete wavelet transform (DWT), revised principal component analysis (RPCA) and auto-regression	Support vector machine (SVM)	Accuracy (96.3), sensitivity (25.3)
N. Kulkarni et al. (2016) [28]	Spectral-, wavelet-, and complexity-based features	Support vector machine (SVM)	Accuracy (96.3), sensitivity (85.3), specificity (86.12)

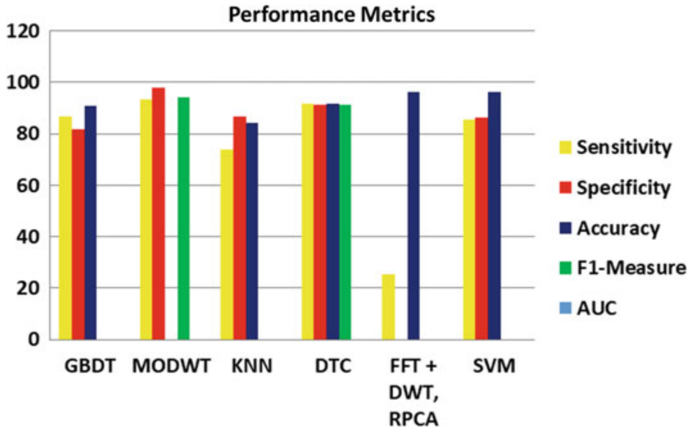


Fig. 4 Performance metrics

Table 3 Various classification classes with database

S. No.	Classifier	Preprocessing	Feature Extraction	Dataset	Classes
1	Decision Tree C4.5 classifier	Yes	Yes	The IRCCS Centro Neurolesi 'Bonino-Pulejo'	Healthy control, Alzheimer's, MCI
2	Principal component analysis (PCA), partial least square (PLS)	Yes	No	Department of Medical Informatics, Institute for Biomedical Engineering, University of Technology Graz	classification of the subject feedback session
3	Support vector machine (SVM), K-nearest neighbor (KNN)	Yes	Yes	Sina and Nour Hospitals, Isfahan, Iran	MCI
4	Support vector machine (SVM)	Yes	No	–	Healthy control, Alzheimer's
5	Convolutional neural networks (CNN)	Yes	Yes	Alzheimer Disease Neuroimaging Initiative (ADNI)	Alzheimer's, MCI
6	Multilayer perceptron (MLP), logistic regression (LR) and support vector machine (SVM)	Yes	Yes	IRCCS Centro Neurolesi Bonino-Pulejo of Messina (Italy)	Healthy control, Alzheimer's, MCI
7	Feature selection, regression, support vector machine (SVM) and ensemble	Yes	Yes	Clinicas Hospital School of Medicine (FMUSP-HC), University of Sao Paulo	Alzheimer's
8	K-nearest neighbor (KNN)	Yes	Yes	IRCCS Centro Neurolesi Bonino-Pulejo	Healthy control, Alzheimer's
9	Principal component analysis (PCA)	Yes	Yes	https://www.epileptologie-bonn.de/cms/front_content.php?idcat=193	Epilepsy
10	Linear discriminant analysis (LDA)	Yes	Yes	Hospital of Sun Yat-sen University (SYSU) in Guangzhou (China)	Normal elderly control, Alzheimer's

References

1. Mattson M (2004) Pathways towards and away from Alzheimer's disease. Nature 430
2. May P, Ehrlich HC, Steinke T (2006) ZIB structure prediction pipeline: composing a complex biological workflow through web services. In: Nagel WE, Walter WV, Lehner W (eds) Euro-Par 2006. LNCS, vol 4128, pp 1148–1158. Springer, Heidelberg (2006); Meek PD, McKeithan K, Shumock GT (1998) Economics considerations of Alzheimer's disease. Pharmacotherapy 18(2 Pt 2):68–73
3. McKhann GM et al (2011) The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's Dementia 7(3):263–269
4. Greicius MD et al (2004) Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI. Proc Natl Acad Sci 101(13):4637–4642
5. Stam CJ (2005) Nonlinear dynamical analysis of EEG and MEG: review of an emerging field. Clin Neurophysiol 116(10):2266–2301
6. Jeong J (2004) EEG dynamics in patients with Alzheimer's disease. Clin Neurophysiol 115(7):1490–1505
7. Berger H (1929) Uber das elektrenkephalogramm des menschen. Archiv fur Psychiatrie und Nervenkrankheiten 87(1):527–570
8. Webster JG (1998) Medical instrumentation: application and design, 3rd edn. Wiley & Sons
9. Joy J, Peter S et al (2013) Denoising using soft thresholding. Int J Adv Res Electr Electron Instr Eng 2(3):1027–1031
10. Walters-Williams J, Li Y (2012) BMICA independent component analysis based on B-spline mutual information estimation for EEG signals. Can J Biomed Eng Technol 3(4):63–79
11. Prasad VVKDV et al (2014) A new wavelet packet based method for denoising of biological signals. Int J Res Comput Commun Technol 2(10):1056–1062; Kaur S (2014) IJECS 3(8):7965–7973, p 7973
12. Fiscon G, Weitschek E, De Cola MC, Felici G, Bertolazzi P (2018) An integrated approach based on EEG signals processing combined with supervised methods to classify Alzheimer's disease patients. In: 2018 IEEE international conference on bioinformatics and biomedicine (BIBM), Madrid, Spain
13. Khan AT, Afrin S, Rahman T (2018) Comparison of principal component analysis and partial least square discriminant analysis in the classification of EEG signals. In: 2018 IEEE international workshop on signal processing systems (SiPS), Cape Town, South Africa, pp 1–6
14. Siuly S et al (2020) A new framework for automatic detection of patients with mild cognitive impairment using resting-state EEG signals. IEEE Trans Neural Syst Rehabil Eng 28(9):1966–1976
15. Nobukawa S et al (2020) Classification methods based on complexity and synchronization of electroencephalography signals in Alzheimer's disease. Front Psychiatry 11
16. Ismail M, Hofmann K, El Ghany MAA (2019) Early diagnoses of alzheimer using EEG data and deep neural networks classification. In: 2019 IEEE global conference on ınternet of things (GCIoT), Dubai, United Arab Emirates
17. Gonzalez HA, Yoo J, Elfadel IM (2019) EEG-based emotion detection using unsupervised transfer learning. In: 2019 41st Annual ınternational conference of the IEEE engineering in medicine and biology society (EMBC). IEEE
18. Ieracitano C et al (2019) A time-frequency based machine learning system for brain states classification via eeg signal processing. In: 2019 International joint conference on neural networks (IJCNN). IEEE
19. Tavares G et al (2019) Improvement in the automatic classification of Alzheimer's disease using EEG after feature selection. In: 2019 IEEE international conference on systems, man and cybernetics (SMC). IEEE
20. DEL R. MILLÁN JOSÉ, et al (2008) Non-invasive brain-machine interaction. Int J Pattern Recogn Artif Intell 22(5):959–972

21. https://towardsdatascience.com/independent-component-analysis-ica-in-pythona0ef0d b0955e
22. Motamedi-Fakhr S et al (2014) Signal processing techniques applied to human sleep EEG signals—a review. Biomed Signal Process Control 10:21–33
23. Albaqami H et al (2020) Automatic detection of abnormal EEG signals using wavelet feature extraction and gradient boosting decision tree. arXiv preprint arXiv:2012.10034
24. Ge Q et al (2020) A robust discriminant framework based on functional biomarkers of EEG and ıts potential for diagnosis of alzheimer's disease. Healthcare 8(4). Multidisciplinary Digital Publishing Institute
25. Durongbhan P et al (2019) A dementia classification framework using frequency and time-frequency features based on EEG signals. IEEE Trans Neural Syst Rehab Eng 27(5):826–835
26. Fiscon G et al (2018) Combining EEG signal processing with supervised methods for Alzheimer's patients classification. BMC Med İnf Dec Making 18(1):1–10
27. Deepa R, Shanmugam A, Tamilselvan E (2017) EEG feature extraction and classification of alzheimer's disease using support vector machine classifier
28. Kulkarni NN, Bairagi VK (2017) Extracting salient features for EEG-based diagnosis of Alzheimer's disease using a support vector machine classifier. IETE J Res 63(1):11–22

Computer Vision Human Activity Recognition Using Cumulative Difference Energy Representation-Based Features and Employing Machine Learning Techniques

R. Sathya and E. Gokulakannan

Abstract Vision-based human action and motion recognition system are very demanding assignments for many computer processor request comprises criminal investigations, public video surveillance system, and applications of sports. For lengthy videos capture, it is very hard to identify within a video frame for a particular human action. This paper presents an efficient technique to human action detection, feature extraction, and recognition. In the first phase, in order to identify the moving object, the cumulative difference scheme is applied. The next phase, in aim feature tracking based on the proposed cumulative difference energy representation (CDER) technique is used. In our proposed research, we have evaluated different machine learning techniques methods, namely K-nearest neighbors, support vector machine, random forest, naïve Bayes, and decision tree. Experimental results were conducted on the publically available human action KTH datasets considering six activities, viz. (hand clapping, walking, hand waving boxing jogging, and running). The research results express that proposed CDER technique outperforms the majority of previous schemes, accomplishing accuracy of 98.67% for SVM classifier.

Keywords Frame differencing · KTH dataset · Decision tree · KNN · Random forest · Naïve Bayes · SVM · Action recognition

1 Introduction

Individual person action classification has involved important consideration in the region of computer vision society in current decades and has encouraged the growth of a large diversity of applications, including video archive indexing, human–computer vision communication, and finally video surveillance monitoring. Image and video

R. Sathya (✉)
Kongunadu College of Engineering & Technology (Autonomous), Trichy, India
e-mail: rsathyamephd@gmail.com

E. Gokulakannan
MRK Institute of Technology, Cuddalore, India

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_40

surveillance systems are rapidly growing and used for day-by-day protection. Surveillance cameras are offered with a variety of volumes. The image speed of every camera varies based on the camera volumes or resolution. Typically, the image velocity for a surveillance camera of standard feature (1280 × 720) is approximately thirty fps. In a computerized video supervision system, dynamic object identification, tracking, and recognizing of moving conditions leftover a difficult mission.

Perfect techniques modeling of motion classification must challenge with a huge number of disputes. In current years, investigating has moved on classifying variety of human actions using uncomplicated feature vector. Human being actions can be classified into four stages according to Aggarwal et al. [1, 2]. They point out that actions can be categorized as human gestures, human interactions, human actions and people group behaviors, distinct as follows:

Human Gestures: Essential actions of a human's body element (signaling a hand or lifting a leg). Human Interactions: Human behaviors that involve more than one people (hand shaking, kicking, fighting, and hugging). Human Actions: only one-person behaviors. These can be performed of numerous actions (walking, running, clapping, and waving). People Group Actions: Actions presented by grouping of several persons (more than one group of people fighting, variety of persons dancing in part hall). For the intention of transparency, we describe human movements and behaviors for the rest of the research paper as follows: Human movements: a set of transitional positions/sub-states that collectively explain the action of the person {i.e., jumping jacks, run, skip} Human behaviors: a variety of chronological activities {walk, jump, run}.

This research paper is presented as follows. Upcoming section presents a succinct explanation of the related work from the previous publication. Section 3 presents the KTH human action dataset. Section 4 explains the proposed cumulative difference energy representation (CDER) for human action recognition. Section 5 describes the machine learning techniques for human action classification. Section 6 explains the numerical experimental results and discussion along with the estimation of performance analysis. Last of all, the conclusion and future work is presented.

2 Related Work

Several techniques are newly proposed for moving human action identification and classification from video sequences. The new methods employ a variety of methods to imprison person action in sequence such as background subtraction [3], optical flow [4], and frame difference [4]. In accumulation, various types of feature extraction methods are employed such as HOG [5], SIFT descriptor [6] and even though scale-invariant feature transform and histograms of oriented gradients systems accomplish important evolution in several classification missions, they cannot be measured as a worldwide extractor of feature for resolve all missions [6].

Now years millions of human action videos are taken and uploaded to a lot of application, different action and diverse action are done by the single person, besides that various people can do the same action in diverse action way, that's create it in actuality an enormous challenge to classify the human action. In computer vision, technology presents many publically available datasets. Abdelbaky et al. [7] and Elharrouss et al. [8] are focused on KTH dataset for human action recognition system.

Cao et al. [9] projected a MAPF for multifaceted motions to conquer the unexpected movement and affine conversion troubles in motion classification. Zhao et al. [10] concentrated on a few problem connected to illustration classification, such as pose changes, action shadow, parallel coloring allocation in the static background image, overlapping and enlightenment transforms. The foreground or background subtraction [11, 12] was the early stage in moving object and hand gesture segmentation. To face local optimization problems [13, 14], a biased search-based method was employed for distillation of the positioned person intention in multifaceted background scenarios.

3 Dataset

The employee of openly accessible databases permits for the evaluation of various methods and provides imminent into the inabilities of relevant techniques. The proposed research explains the very used person activity database.

3.1 KTH Human Action Dataset

The KTH Royal Institute of Technology produced this benchmark human action database [15] in two thousand four accomplishing a significant objective in the video processing society. At that time, this benchmark video developed the peak video benchmark dataset with series of person activities in excess of various environments. The KTH human motion database includes six varieties of person activities (running, boxing, hand clapping, walking, hand waving, and jogging) executed by twenty-five different actors. Four diverse circumstances are used: outdoors (s1), outdoors with scale variation (s2), outdoors with different clothes (s3), and indoors with lighting variation (s4). There are an overall of $25 \times 6 \times 4 = 600$ video records for both combination of twenty-five actors, six actions, and four scenarios.

4 Cumulative Difference Energy Representation (CDER) for Human Action Recognition

The CDER representation is created on the preprocessed human action sequences in KTH. In this section, we initially establish the preprocessing processes and secondly

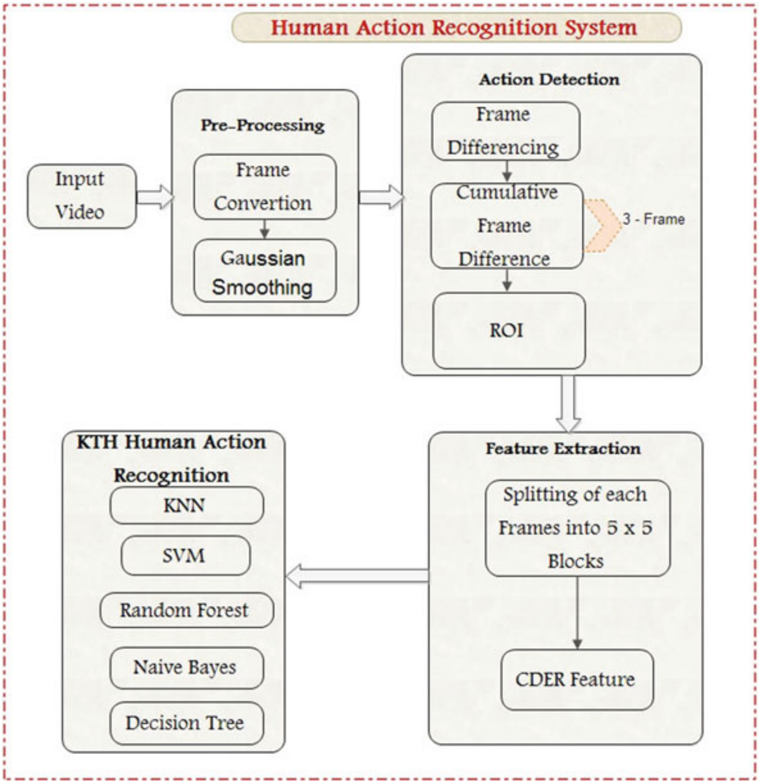


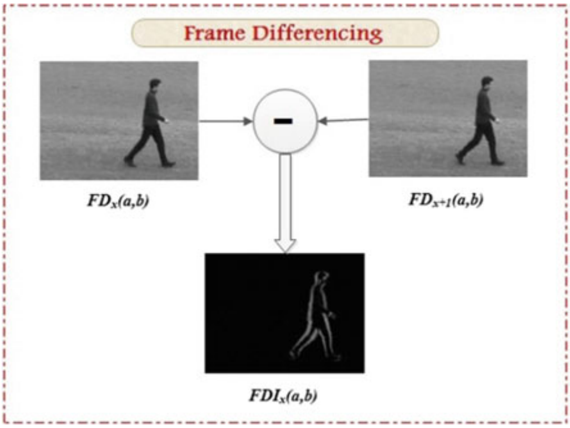
Fig. 1 Architecture for proposed CDER system

propose how to detect motion using frame differencing. Next, we propose a new method of cumulative difference and finally extract the features from cumulative difference images using proposed CDER feature extraction algorithm. The proposed CDER recognition system is shown in Fig. 1.

4.1 Preprocessing

As declared previously that's we depend on human action KTH video database to classify the human motion, initial stage in preprocessing is converting the KTH video to frames to work on each and every frame separated. The KTH video is processed at twenty-five fps. Smoothing of image is made by Gaussian convolution techniques with 3×3 size of a kernel and variance $\sigma = 0.6$. This method is most important to preprocess KTH person video sequences to eliminate noise for further tracking and recognition.

Fig. 2 Frame differencing



4.2 Motion Detection Using Frame Difference

Images are preferred based on the intensity deviation between background images and foreground images. The absolute difference frame achieved by subtracting the present image is at time x with successive image at time $x + 1$ on a pixel-by-pixel origin. For obtaining foreground image, the pixel range superior than the threshold Th_n is position to 1, whereas those lesser than the threshold Th_n are assigned to zero as background image pixels, where frame difference image (FDI) is indicated by FDIx(q,r) and walking action FDI image is shown in Fig. 2.

$$\mathrm{FDI}_x(q,r) = \begin{cases} 1, & |\mathrm{FD}_{\mathrm{x}}(q,r) - \mathrm{FD}_{x+1}(q,r)| > \mathrm{th}_n; \\ 0, & |\mathrm{FD}_{\mathrm{x}}(q,r) - \mathrm{FD}_{x+1}(q,r)| > \mathrm{th}_n; \end{cases} \tag{1}$$

$$\begin{gathered} \mathrm{FD}_x(q,r) = |\mathrm{FD}_x(q,r) - \mathrm{FFD}_{x+1}(q,r)| \\ 1 \le a \le w,\ 1 \le b \le h \end{gathered} \tag{2}$$

where $\mathrm{FD}_x(a, b)$ is current frame and $\mathrm{FD}_{x+1}(q, r)$ is successive image. $\mathrm{FD}_x(q, r)$ is the intensity of the pixel (q, r) in the xth image, h and w are the height and width of the frame, respectively. Moving human actions can be identified extremely fast by using frame difference method [16].

4.3 Proposed Target Action Detection Using Cumulative Difference

The proposed cumulative differencing technique is applied for predicting the region performance highest intensity. In the proposed three-frame cumulative difference method, the cumulative difference is taken between current frame and successive frame, which gives an outline of the foreground object. Novel four-frame cumulative difference images are computed as follows:

Where $k = 1$, FD_k is the normal frame difference image achieved by subtracting two consecutive frames f_k and f_{k+1}.

$$\mathrm{FD}_k(q,r) = \begin{cases} 1 & f_k(q,r) - f_{k+1}(q,r) > th_n; \\ 0 & f_k(q,r) - f_{k+1}(q,r) < th_n; \end{cases}$$

$$\mathrm{FD}_{k+1}(q,r) = \begin{cases} 1 & f_{k+1}(q,r) - f_{k+2}(q,r) > th_n; \\ 0 & f_{k+1}(q,r) - f_{k+2}(q,r) < th_n; \end{cases} \quad 3$$

$$\mathrm{FD}_{k+2}(q,r) = \begin{cases} 1 & f_{k+2}(q,r) - f_{k+3}(q,r) > th_n; \\ 0 & f_{k+2}(q,r) - f_{k+3}(q,r) < th_n; \end{cases}$$

$$\mathrm{CFD}_k(q,r) = \mathrm{FD}_k(q,r) + \mathrm{FD}_{k+1}(q,r) + \mathrm{FD}_{k+2}(q,r) \tag{4}$$

4.4 Proposed CDER Feature Extraction Algorithm in KTH Database

Human motion recognition is based on two varieties of information: shape motion information and dynamic motion information. Human shape motion information refers to the human appearance, such as body width, height, body part proportions, hunching, hair, dressing, and so on. Several human shape motion information remains inflexible or related during human walking. Human motion information extracted from a distinct image belongs to shape motion information. The most important motivation of the novel proposed cumulative difference energy representation (CDER) is to calculate the average of pixel value from the proposed cumulative frame difference image.

The identified ROI human motion information is extracted. The extracted ROI region is considered for CDER feature extraction procedure. The extracted ROI images are various sizes. The extracted objects are resized without any significant information loss for further analysis. Since this work considers the extracted ROI image of size 60 × 40 for the extraction of CDER features, the resized extracted region is separated into $s \times t$ blocks, where s is the height of the ROI region and t is the width of the ROI region. The ranges of s and t are fixed in such a way that each blocks are same size. Where $s = 5$ and $t = 5$ since this proposed research considers the each block of size 12 × 8. The calculated intensity amount of pixels in each block is called cumulative difference energy representation (CDER). The proposed technology takes out 25-dimensional CDER features from three-frame cumulative frame difference image.

5 Machine Learning Techniques and Discussion

Classification approaches under supervised learning procedure in order to calculate known initialize the information to an assured class label. The following six human actions were measured for action recognition: hand waving, walking, running, boxing, jogging, and hand clapping. For gathering the KTH database, each actors was asked to carry out each action several times in face of the static camera indoor and outdoor scenario. Many of the methods typically utilized classification methods are explained as below.

5.1 K-nearest Neighbors (KNN) Classifier

The KNN techniques normally choose k-closest models from the training attributes to the original case, and the class with the maximum amount of votes is allotted to the assessment case.

Procedure for K-nearest neighbor techniques

- Obtain the input feature vector
- Select starting rate of k
- Separate the training feature data and testing feature vector
- For accomplishing essential intention iteration for every training vector position
- Locate the space between analyze feature vector and every row of training feature vector. (Euclidean Distance approach is the excellent method)
- Assemble the considered space in climbing order based on space vectors
- Select the highest k range from arranged vector
- Analyze the highest label model
- Get the aim category.

5.2 Support Vector Machine (SVM) Classifier

SVM [16] is powerful techniques for data reorganization. One of the easy and valuable techniques in supervised method is SVM techniques. SVM can be structured using a variety of kernel function to improve the accuracy: polynomial, Gaussian, and Sigmoidal. SVM is well appropriate for mutually unstructured and structured feature data.

5.3 Random Forest Classifier

A random forest is a compilation of decision trees prepared with random features [17]. Random forests work as follows.

- Choose parameter T, which is the numeral of trees to grow.
- When growing each tree, the following is carried out.

5.4 Naïve Bayes (NB) Classifier

Naive Bayes (NB) technique [18] is one of the best techniques with strapping autonomous theory between the extracted vectors.

$$a = \text{avrg}\max_{a} p_1(a)\prod_{n=1}^{k} p_1\left(\frac{b_n}{a}\right)$$

P_1 (a) is called class probability and is provisional probability.

5.5 Decision Tree (J48) Classifier

Decision tree [19] develops tree perceptions to locate the answer of the different actions. The information index for an exacting feature data Z at a node is calculated as

$$\text{Information Index }(X, Z) = \text{ Entropy}_1(X) - \sum_{\text{Value at } z} \frac{|X|}{|X_n|}\text{Entropy}_1(X)$$

where X is the combination of instance at that exacting node and

$$|X| : \text{ Cordinality}_1$$

$$\text{Entropy}_1(X) = \sum_{n-1}^{X} -p_n \log_2 p_i$$

6 Numerical Experimental Analysis and Results

In this section, we discuss the numerical experimental results of proposed CDER feature set for various machine learning techniques. The proposed experiments are carried out in KTH dataset. To demonstrate the procedure, concepts, and human motion recognition based on machine learning techniques, we carried out experiments on every frame of various human actions in different scenarios.

6.1 Classification Results Using Proposed CDER Features with Various Machine Learning Techniques

KTH human action dataset was used in this work and is one of the major person action video database; it contains six human activity classes. As stated above, leave-one-person-out experimental system is utilized in the proposed research, where every run utilizes 20 actors for training and five actors for testing.

In this subsection, we represent the performance of proposed CDER feature space for different machine learning classifiers. Figure 3 represents the confusion matrix of various machine learning techniques. Figure 3 Table (a) concentrates the confusion matrices of KTH database using KNN classifier. Figure 3 Table (b) shows the confusion matrices of the SVM classifier with RBF kernel for the CDER feature. Figure 3 Table (c) illustrates the confusion matrix for KTH action recognition with random forest classifier. Figure Table (d) presents the confusion matrices of action recognition using naïve Bayes classifier. Figure 3 Table (e) illustrates the confusion matrix for KTH action recognition with decision tree classifier.

Precision and recall obtained with different machine learning classifiers are shown in Fig. 4. Figure 5 presents a comparison between the overall human action recognition accuracy achieved using proposed cumulative difference energy representation (CDER) feature space with various classifiers, as shown the proposed CDER feature technique proved higher accuracy on the action recognition performance using SVM classifier. The action recognition accuracy obtained for KNN, SVM, random forest, naïve Bayes, and decision tree classifiers were 97.3, 98.67, 96.8, 97.6, and 98.1, respectively. The highest recognition rate obtained for the SVM RBF kernel was 98.67% on proposed CDER features for KTH actions.

6.2 Comparison

Table 1 presents a similarity between the proposed techniques and a collection of further earlier proposed methods that employ leave-one-out system. The experimental analysis demonstrates that for the benchmark KTH human action database our experimental analysis is the most excellent of the previous proposed system.

	Walking	Running	Jogging	Hand Waving	Hand Clapping	Boxing
Walking	97.9	1.2	0.3	0.3	0	0.3
Running	1.5	97.8	0.2	0	0	0.5
Jogging	0	2.1	97.1	0.5	0.3	0
Hand Waving	0	0	1.3	96.8	1.9	0
Hand Clapping	0	0	0.7	1.4	97.9	0
Boxing	0	0	0	1.6	2.1	96.3

(a) KNN

	Walking	Running	Jogging	Hand Waving	Hand Clapping	Boxing
Walking	98.8	0.8	0.2	0.1	0	0.1
Running	1.3	98.1	0	0	0	0.6
Jogging	0	1.5	98.3	0	0.2	0
Hand Waving	0.2	0	0.5	98.7	0.6	0
Hand Clapping	0	0.2	0.2	0.7	98.9	0
Boxing	0.3	0	0	0.4	0.1	99.2

(b) SVM

	Walking	Running	Jogging	Hand Waving	Hand Clapping	Boxing
Walking	96.9	1.7	0.7	0	0.6	0.1
Running	1.9	96.8	0.5	0.5	0	0.3
Jogging	0	1.8	97.5	0.7	0	0
Hand Waving	0.4	0	1.6	95.8	0.7	1.5
Hand Clapping	0.2	0	1.2	2.1	96.5	0
Boxing	0	0.4	0	0.5	1.8	97.3

(c) Random Forest

	Walking	Running	Jogging	Hand Waving	Hand Clapping	Boxing
Walking	98.1	0.9	0.2	0.5	0	0.3
Running	0	97.9	0.7	1.4	0	0
Jogging	0	1.1	98.5	0.1	0.3	0
Hand Waving	0	0	1.3	96.8	1.9	0
Hand Clapping	0	0.7	1.1	0.5	96.5	1.2
Boxing	1.5	0.5	0.2	0	0	97.8

(d) Naïve Bayes

	Walking	Running	Jogging	Hand Waving	Hand Clapping	Boxing
Walking	98.5	0.8	0.2	0	0	0.5
Running	1.5	98.1	0	0	0.2	0.2
Jogging	0	1.2	98.3	0.3	0	0.2
Hand Waving	1.2	0	0.6	97.3	0.9	0
Hand Clapping	0	0.4	0.7	0.4	98.5	0
Boxing	0.1	0	0.1	0.6	1.3	97.9

(e) Decision Tree

Fig. 3 Confusion matrix of various machine learning techniques

7 Conclusion and Future Work

This proposed research paper presents the best techniques for person activity classification based on the grouping of cumulative difference energy representation (CDER) frames. The proposed system is utilized three stages: detection of human actions

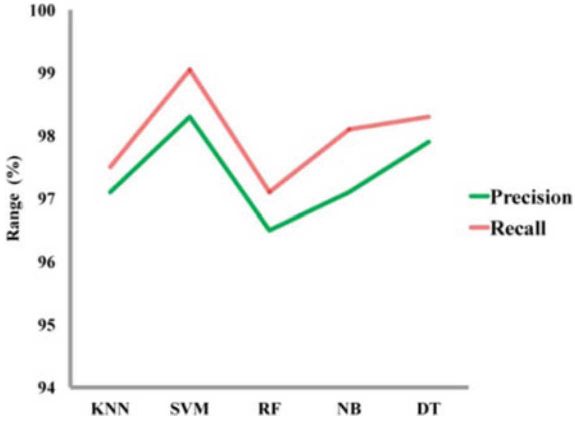


Fig. 4 Precision and recall value for different classifier using KTH dataset

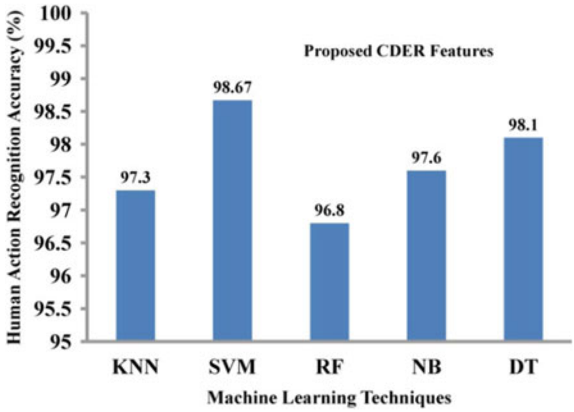


Fig. 5 Comparison for proposed CDER feature with machine learning techniques

Table 1 Comparison between earlier system and proposed system

Method	KTH recognition accuracy (%)
Our proposed method	98.67
Moussa et al. [20]	97.89
Jin et al. [21]	96
Yong et al. [22]	96
El-Henawy et al. [23]	95
Liu et al. 2017 [24]	94
Kaminski et al. [25]	92
Dasari et al. [26]	87

positions, features description, and action classification. For the first step, cumulative difference technique is used, and in second step, a novel cumulative difference energy representation is used for feature extraction. Finally, the traditional machine learning techniques were utilized for KTH human action classification. The proposed recognition system achieved the wonderful recognition rate for SVM classifier (98.67%) which is comparable with the proposed other machine learning classifier.

Future work contains relating the proposed CDER approach on various environment databases, such as Weizmann, sports, UCF dataset, and real-time activity datasets. These types of data base are very difficult than the ones used here and the recognition approach may well require a quantity of development to accomplish acceptable action classification rate and the utilize of a video sequence of diverse human activity to division it then classifies every human activity is an additional new method of investigating activity in the further work.

References

1. Aggarwal, J., Cai, Q., 1997. Human motion analysis: A review. In: Proc. Nonrigid and Articulated Motion Workshop. IEEE, pp. 90–102.
2. Aggarwal, J., Ryoo, M., 2011. Human activity analysis: A review. ACM Comput. Surv., 0360–0300 43, 16:1–16:43. http://dx.doi.org/https://doi.org/10.1145/1922649.1922653
3. Ray, K. S. & Chakraborty, S. (2017a). An efficient approach for object detection and tracking of objects in a video with variable background, arXiv preprint [Online]. Available at: <http:// arxiv.org/ abs/1706.02672>.
4. Ahad MAR, Tan JK, Kim H, Ishikawa S (2009) Temporal motion recognition and segmentation approach. Int J Imaging Syst Technol 19(2):91–99
5. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: IEEE computer society conference on computer vision and pattern recognition, 2005. CVPR 2005. IEEE, vol 1, pp 886–893.
6. Mikolajczyk K, Schmid C (2005) A performance evaluation of local descriptors. IEEE Trans Pattern Anal Mach Intell 27(10):1615–1630
7. Abdelbaky A, Aly S (2020) Human action recognition using short-time motion energy template images and PCANet features. Neural Comput Appl. Springer Nature. https://doi.org/10.1007/s00521-020-04712-1
8. Elharrouss O, Almaadeed N, Al-Maadeed S, Bouridane A, Beghdadi A (2020) A combined multiple action recognition and summarization for surveillance video sequences. Appl Intell. https://doi.org/10.1007/s10489-020-01823-z
9. Cao S, Wang X, Xiang K (2017) Visual object tracking based on Motion-Adaptive Particle Filter under complex dynamics. EURASIP J Image Video Process 2017(1):1–21. https://doi.org/10.1186/s13640-017-0223-0
10. Zhao H, Xiang K, Cao S, Wang X (2016) Robust visual tracking via CAMShift and structural local sparse appearance model. J Vis Commun Image Represent 34:176–186. https://doi.org/10.1016/j.jvcir.2015.11.008
11. Thabet E, Khalid F, Sulaiman PS, Yaakob R (2021) Algorithm of local features fusion and modified covariance-matrix technique for hand motion position estimation and hand gesture trajectory tracking approach. Multimedia Tools Appl 80(4):5287–5318. https://doi.org/10.1007/s11042-020-09903-5
12. Nallasivam M, Senniappan V (2021) Moving Human Target Detection and Tracking in Video Frames. Studies in Informatics and Control 30(1):119–129

13. Sultana M, Mahmood A, Jung SK (2020) Unsupervised moving object detection in complex scenes using adversarial regularizations. IEEE Trans Multimedia. https://doi.org/10.1109/TMM.2020.3006419
14. Sargano AB, Angelov P, Habib Z (2017) A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition. Appl Sci 7(1):110
15. Laptev I, Caputo B (2011) Recognition of human actions
16. Sathya R, Kalaiselvi Geetha M (2013) Vision based traffic police hand signal recognition in surveillance video—a survey. Int J Comput Appl (IJCA) 81(9):1–10
17. Breiman L (2001) Random forests. Mach Learn 45:5–32
18. Farid DM, Rahman MZ, Rahman CM (2011) Adaptive intrusion detection based on boosting and nave Bayesian classifier. Int J Comput Appl 24:12–19
19. Quinlan JR (1986) Introduction of decision trees. Mach Learn 1:81–106
20. Moussa MM, Hamayed E, Fayek MB, El Nemr HA (2015) An enhanced method for human action recognition. J Adv Res 6:163–169
21. Jin C-B, Li S, Kim H (2017) Real-time action detection in video surveillance using sub-action descriptor with multi-CNN. arXiv:1710.03383
22. Yang J, Ma Z, Xie M (2015) Action recognition based on multiscale oriented neighborhood features. Int J Sig Process Image Process Pattern Recognition 8(1):241–254
23. El-Henawy I, Ahmed K, Mahmoud H (2018) Action recognition using fast HOG3D of integral videos and Smith-Waterman partial matching. IET Image Process 12(6):896–908
24. Liu H et al (2017) Study of human action recognition based on improved spatio-temporal features. In: Human motion sensing and recognition. Springer, pp 233–250
25. Kaminski Ł, Mackowiak S, Domanski M (2017) Human activity recognition using standard descriptors of MPEG CDVS. In: 2017 IEEE international conference on multimedia & Expo Workshops (ICMEW). IEEE
26. Dasari R, Chen CW (2018) MPEG CDVS feature trajectories for action recognition in videos. In: 2018 IEEE conference on multimedia information processing and retrieval (MIPR). IEEE

A Survey on Advancements of Real-Time Analytics Architecture Components

Rajnish Dashora and M. Rajasekhara Babu

Abstract With the evolving technological landscape, Real-Time Analytics has become a de-facto practice and area of attraction for various industrial applications. Real-Time Analytics has proven to be impactful in plenty of critical problem statements, and a wide range of use cases have been raised in the real-time world. Working with streaming data includes various steps and aspects of software development and engineering starting from data collection to movement to processing and building actionable insights. Every such component of streaming analytics architecture needs to be designed and developed keeping in mind that the materialization of insights generated with this kind of dataset needs to be nearly real time. Many industrial organizations and large-scale enterprises have grown by adopting and investing heavily into building their systems using event-based streaming architectures. In this article, we intend to provide insights on primarily three aspects of architectural components in the context of Real-Time Analytics as well as advancements in the field. These aspects include recent research and development with regards to Real-Time Analytics Architecture and its use cases, industrial applications, development of tools and technologies. We also attempt to expose some open challenges and research issues that need further attention from researchers and industrial experts.

Keywords Big Data · Real-Time Analytics · Data processing · Data flow · Streaming · Streaming Data Architectures

R. Dashora (✉) · M. R. Babu
School of Computer Science and Engineering, VIT University, Vellore 632014, India
e-mail: rajnish.dashora2016@vitstudent.ac.in

M. R. Babu
e-mail: mrajasekharababu@vit.ac.in

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_41

1 Introduction

In pursuit of real-world problem solving and building the world a better place to live, researchers and practitioners have been consistently investing in the evolution of technology and engineering. It has been vital to model and understand the continuous streams of real-world data and develop solutions to gather, clean, analyze, and deliver insights with such raw and unstructured datasets. Designing large data-intensive applications which can handle streaming data throughout different stages of data transformation and flow helps enterprises design better systems. With reliable architecture for Real-Time Analytics, the robustness of such Big Data Systems can be assured.

With this survey, we want to share state-of-the-art technology and research advancements alongside industrial developments happening across the world with regards to Real-Time Analytics. The article also discusses trends in fields such as transportation, smart cities, health care, battlefields, etc.

1.1 Related Surveys on Real-Time Analytics Architecture Components

Each stage of Big Data System starts from data ingestion, data loading, and processing, information extraction, data cleaning, Big Data integration, data analysis, data loading and transformation, data visualization, data storage, metadata management, data lifecycle management, data security, and privacy management has made an impact on how to design and architect software for these use cases and this has caused a paradigm shift in software engineering [1]. A survey article, "Big Data Systems: A Software Engineering Perspective" details out the above emphasizing software engineering aspects [1].

Another recent article with the title "Mapping the Big Data Landscape: Technologies, Platforms, and Paradigms for Real-Time Analytics of Data Streams" is a thorough read. It talks about the "4Vs" of Big Data, velocity, volume, variety, and value/veracity. "4Vs" represent various aspects of information that can be derived such as dimensionality, latency, and throughput. Task distribution and real-time scheduling, fault tolerance with regards to data preservation, message delivery, and task preservation, data availability, computational paradigms and architectures such as MapReduce, Lambda and Kappa architectures, and various message passing systems have developed with advancements in Real-Time Analytics [2].

Also, the article, "A Survey on Stream-Based Recommender Systems" elaborates on how Real-Time Analytics and stream-based architectures help build better recommender systems [3]. In "The Programmable Data Plane: Abstractions, Architectures, Algorithms, and Applications," Oliver et al. [4] presents recent advancements in enabling programmability of data planes and using programmable data planes for various use cases.

1.2 Research Questions and Organization of the Article

With this article, we intend to address the below questions in favor of collating knowledge around recent trends in Real-Time Analytics and contribute back to the community of researchers and practitioners by sharing highlights and insights on development in this area,

Question 1. What are the recent advancements and research around Real-Time Analytics Architecture components?

Question 2. What are the recent industrial impacts with advancements in Real-Time Analytics Architecture components?

Further, the article is organized primarily around these two questions. First, we discuss academic perspectives on the advancements in Real-Time Analytics Architecture components where we deep-dive in the kind of research and developments that are going to support this field. We also discuss the tools and technologies getting developed in each of the different phases and components. Here, the Real-Time Analytics Architecture components are categorized based on the following phases namely collections, data flow, processing, storage, and delivery [5]. In the next section, we discuss the industrial and real-world use cases where Real-Time Analytics has been useful. We will also gather context about the kind of research and advancements that have been supporting the evolution of Streaming Data Architecture, hence improving industrial efficiency and performance.

2 Advancements of Real-Time Analytics Architecture Components: Academic Perspective

Real-Time Analytics has been one of the most interesting and fast-paced paradigms for researchers and practitioners. Research publications from various industrial use cases and of different stages and components of Streaming Data Architecture (expressed in Fig. 1) were searched from the past 2 years, were selected to study and review. A summary of such articles is presented in the following sections based on stages and categorization such as data collection, data flow, data processing, data storage, and data delivery. This section also attempts to cover on few open-source projects supporting the development of Real-Time Analytics Architecture components.

2.1 Collection

For data collection, in bulk applications, the collection process is done at the edge servers and is integrated with various data collection algorithms and data serialization.

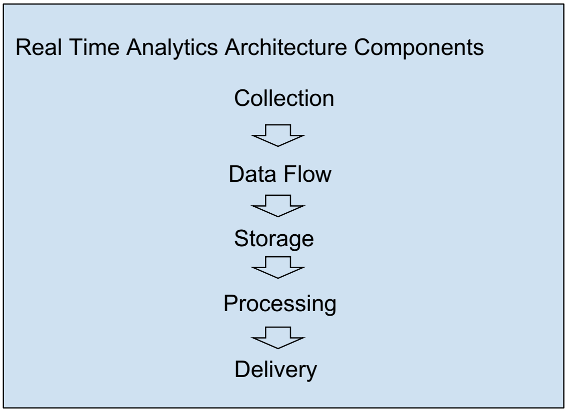


Fig. 1 Real-Time Analytics Architecture components

The collection mechanism directly interacts further with data flow mechanisms. The edge servers are designed and developed based on use cases and specialize in handling the peculiarity of applications. Commonly edge servers or other words sources of streaming data in the real world are IoT sensors collecting and transferring signals and environment data, web application servers sending a variety of logs such as system, server and security logs, click-stream data from mobile applications, websites, etc.

Peng et al. [6] designed data collection algorithms using distributed computing and multi-threading to gain high concurrency and high performance in data collection and data processing for the Internet of Medical Things (IoMT). With a separation of reception and processing of data, there can be significant optimization in resource utilization and processing time while dealing with large-scale distributed data collection challenges. Guan-Chen et al. [7] leveraged advanced deep learning and image processing technologies combined with UAV aerial video, interactive sharing screens, and electronic map interface to implement a microscopic real-time traffic monitoring system for wide area which could achieve the accuracy of traffic flow calculation to nearly 95%. Such a system can be further deployed widely for roadside surveillance in the future. Kokouvi et al. [8] worked on building an adaptive data collection mechanism that reduces period monitoring overhead by 55–70% and yet maintains a good accuracy score. COCO proposed a data collection framework, here when compared with other modern solutions it is lightweight. Sadettin et al. [9], in the context of the COVID-19 environment, worked on developing a Big Data Architecture for real-time streams of videos and surveillance cameras in favor of detecting people, and distance information to build a social distance and object detection, monitoring, and notification system.

Further, the choice of data formats are equally critical, there has been development on formats such as Thrift [10], Protobuf [11], and Apache Avro [12] for the cases where data is fairly structured and the problem space is well-defined, while for unstructured data streams JSON [13] has been a well-known data format which requires a large amount to validation code to ensure the reasonability of data.

2.2 Data Flow

In a large-scale heterogeneous real-time application architecture, data flow systems aim to unify the environment replacing the unstructured, ad-hoc, decentralized communication mechanism. Here are some recent research articles which are relevant to data flow system architectures.

Pedro et al. [14] worked on building Clonos which is a high-availability and fault-tolerance mechanism useful for production-grade systems. Clonos provides consistent local recovery, high availability, and the flexibility of nondeterministic computations. Clonos is assessed to be comparable with the existing fault-tolerance mechanism of Apache Flink with 5–24% overhead in latency and throughput. Ma et al. [15] proposed one of the first frameworks of real-time and error-tolerant MV imputation also called REMAIN, for the poor quality of real-time streaming data which detects and eliminates anomalies from the dataset further learns incrementally and enhances the imputation model. Lesniak et al. [16] introduced a library built for Kafka Streams to enforce event-based constraints explicitly, for developers and demonstrated the advantages to ensure microservice level security using event-based constraints by providing application-level abstractions.

Few systems keeping data flow and stream constraints such as a distributed stream processing framework, Apache Samza [17], a cluster management framework Apache Helix [18], a distributed messaging system, Apache Kafka [19], and a centralized data store, Apache Flume [20] has gained traction from the community of researchers and industrial experts in recent past.

2.3 Processing

Data-stream processing leverages parallel computing principles across distributed network layers with a high error rate and has to deal with challenges such as coordination, transactional processing, partitions and merges developing scatter–gather mechanisms for maintaining performance. There has been a significant amount of work getting done to improvise on existing data processing systems and frameworks such as, Wang et al. [21] developed the design for Kafka's core design to gain correctness guarantees while building streaming data solutions. Read-process-write cycles in Kafka streams rely on replicated logs for storage and inter-processor communication layers. Zhihan et al. [22] leveraged voluntary computing for Big Data processing, which was found effective for the distributed analysis of complex and large-scale networks. Also, an improvisation over the shortest path first algorithm for calculation and verification of parameters was proposed and found to have better performance. Lin et al. [23] presented a novel summarization system that is a unified extractive-abstractive framework embedded with an inconsistency detection module by working on a hierarchical classification problem. Wu et al. [24] leveraged graph neural networks for aggregation of labels comparing with state-of-the-art models, and the proposed framework performed effectively for real-world crowdsourcing datasets, also worked well for modeling latent correlation of tasks. Rozet et al. [25] proposed a framework named Muse leveraging the MatState sharing technique to reduce trend construction and query recomputation overhead for shared aggregation queries by diverse nested Kleene patterns over streaming data. Traub et al. [26] leveraged the stream slicing technique for window aggregation to gain high throughputs at lower latencies and open-sourced an operator for sliding-window aggregation named Scotty for stream processing systems such as Apache Storm [27].

Large-scale data processing solutions such as Apache Flink [28], Apache Spark [29], and Apache Fluo [30], also frameworks like Apache Tez [31] used for arbitrarily complex directed-acyclic graphs (DAGs) processing has been developed to provide reliable and faster data processing systems.

2.4 Storage

Storage for Real-Time Analytics systems needs to meet the performance requirement over atomicity, consistency, isolation, durability (ACID) and allows communication without coupling and access among different components. There are many options available nowadays for storing processed data, the choice of data stores also depends on how the data is to be delivered. Here are a few articles with relevant research to improvise on storage components of a Streaming Data Architecture.

Michael et al. [32] worked on optimization for pattern matching query performance in event stores and developed a method to accelerate such queries using secondary indexes. Zheng et al. [33] developed a new approach to computational storage called DeltaFS Indexed Massive Directories (IMDs) which leverages available resources such as compute, memory, and network for data computations at application nodes. Izadpanah et al. [34] proposed a new technique PETRA, to build data structures with properties such as persistence, non-blocking transactions along supporting ACID properties. Srikanth et al. [35] leveraged the on-chip caches subsystem and introduced accelerators to gain performance of sparse workloads, hence proposing SortCache as a new processor-centric approach. Singh et al. [36] worked on designing a comprehensive microarchitecture SPX64, to address challenges with traditional scratchpad in favor of generality, security, and performance. Maass et al. [37] built an eventual translation lookaside buffer coherence schemes to improvise data center performance for disaggregated data use cases.

Some popular projects for solving data storage use cases and challenges are being developed and used by industrial production-grade systems as well as by researchers are Apache CouchDB [38], Apache CarbonData [39], Apache Phoenix [40], and Apache Lucene-core [41].

2.5 Delivery

After collecting, processing, and storing the data streams, delivery plays a c crucial role in creating value for the end-users and systems. Most often and common delivery mechanisms are web-based interfaces, mobile applications, and query systems making it easy to access, retrieve, and analyze information to gain insights and affect the real world.

Paschalides et al. [42] leveraged NLP and ML technologies to develop a big data processing systems named MANDOLA which is capable of monitoring detecting, visualizing, and reporting the online spread and penetration of hate speech. Farhat et al. [43] leveraged watermark analysis and prioritization of execution and designed a framework, Klink, to perform process-aware scheduling with optimization on query execution with window operators for event streams minimizing output latency. Li et al. [44] worked on building a framework for time-series forecasting named HyDCNN which is trainable and can capture nonlinear dynamics and linear dependencies for various patterns such as sequential and periodic. Savva et al. [45] leveraged machine learning and extraction of knowledge from historical queries to expedite the data analytics process to provide explanations for aggregate queries for exploratory analysis.

Apache Zeppelin [46], Grafana [47], and Prometheus [48] are some of the open-source solutions for establishing an end-to-end delivery mechanism and interfaces for production-grade large-scale data streaming systems and general-purpose data processing (Table 1).

3 Advancements of Real-Time Analytics Architecture Components: Industrial Trends

Real-Time Analytics systems have made and continue to make a significant impact on various industrial domains and problem spaces affecting and hence improving human life. In this section, we aim to cover some of the recent articles expressing traction toward this field and real-world examples of how Stream Data Architectures are evolving to help industries grow.

3.1 Transportation

Zang et al. [49] worked on building a general encoder–decoder framework to solve challenges with crowd flow prediction for cities in multiple steps, leveraging deep learning and Real-Time Analytics. Deng et al. [50] proposed a transportation demand forecasting model exploring co-evolving patterns between various modes of transport hence solving for spatio-temporal forecasting. Daghistani et al. [51] introduced

Table 1 Summary on advancements of Real-Time Analytics Architecture components: academic perspective

Real-Time Analytics Architecture component	Trending research	Popular projects
Collection	Data collection algorithms the Internet of Medical Things (IoMT) [6], data collection mechanism that reduces period monitoring [8]	Thrift [10], Protobuf [11], Apache Avro [12], JSON [13]
Data flow	Clonos [14], REMAIN [15]	Apache Samza [17], Apache Helix [18], Apache Kafka [19], Apache Flume [20]
Processing	Kafka's core design to gain correctness guarantees [21], voluntary computing for Big Data processing [22], Muse [25]	Apache Flink [28], Apache Spark [29], and Apache Fluo [30], Apache Tez [31]
Storage	SortCache [35], DeltaFS Indexed Massive Directories (IMDs) [33]	Apache CouchDB [38], Apache CarbonData [39], Apache Phoenix [40], and Apache Lucene-core [41]
Delivery	MANDOLA [42], HyDCNN [44], Klink [43]	Apache Zeppelin [46], Grafana [47], and Prometheus [48]

SWARM and adaptivity protocol for continuous monitoring of workloads of spatially distributed data streaming in a world of GPS-enabled devices. Fu et al. [52] from Uber have presented how Uber has evolved its real-time data infrastructure solving for business-critical use cases driving flexibility, scalability, and generality within Streaming Data Architectures.

3.2 *Health Care*

Badr [53] performed a survey of various image types in space and time for the medical and healthcare fields. In the current circumstances of the world, medical image data analysis happens to be one of the most important and critical needs where such data is getting generated at large scales and in form of millions of images. With the help of such research, CAD systems are being developed to support doctors and medical workers with applications such as early identification and diagnosis of disease, assessment of risks, and support systems for critical decision making. Can et al. [54] leveraged federated deep learning to build a biomedical informatics solution with data collection from wearable IoT devices to monitor heart activity and maintain the privacy of health data for individuals. Yue et al. [55] developed a CAD algorithm termed HE-CLSTM for automated Real-Time Analytics on time-series data of medical and healthcare images to work in a privacy-preserving manner.

3.3 Smart Cities

Piccialli et al. [56] leveraged deep learning techniques to develop a smart parking system which can forecast space occupancy and optimize traffic flow within parking lots. Djenouri et al. [57] proposed various algorithms named DBSCAN-GTO, kNN-GTO, and FS-GTO, leveraging high-performance computing and ensemble learning for trajectory outlier detection problems for smart cities.

3.4 Miscellaneous

Shahid et al. [58] developed IoTNetWar framework for the Internet of Battlefield Things (IoBT) system applying machine learning and mist computing to gain superiority and hold on information and intelligence by providing data processing and analytics at each layer (i.e., mist, fog cloud, and application), further improving the performance with lessor processing time, latency with better reliability. Shen et al. [59] worked on building a system to profile and segment user data creating insights from non-linguistic audio to infer personalized details and recommendations such as personality and gender. Almaslukh et al. [60] developed a framework to create personalized queries based on geo-social information within streaming data in favor of providing near real-time personalized responses with social graphs.

Apart from the ones mentioned above various other industries and use cases have been affected positively with the use of real-time streaming technologies such as IoT, cyber-physical systems, mining and metallurgy, aerospace, e-commerce, and many more.

4 Open Research Issues

So far, we have surveyed advancement and recent trends with Real-Time Architectures and data processing systems. While there has been significant progress and impact of the research and industrial development getting carried out, there are many open questions and challenges which remain. Problems statements such as demand forecasting for transportation, privacy, and security of data, distributed data collection and fault tolerance, spatio-temporal forecasting, data, and platform generality, governance, network optimization, configuration, and coordination, classification of data streams and dynamically adjusting diversity of datasets [61], knowledge preservation and distribution [62], and differential output testing [63], etc., continue to remain areas of interest for researchers and practitioners.

5 Conclusion

While there has been a diverse range of innovation with regards to Real-Time Analytics, there is still infancy with every stage of Real-Time Analytics Architecture components. This survey attempts to provide an overview of specific problem statements in the field. It also provides an overview of state-of-the-art research and industry efforts getting carried out for Streaming Data Architectures its various phase components such as collection, data flow, processing, storage, and delivery. This survey also details out recent trends in industrial applications in the field of transportation, smart cities, health care, battlefields, etc. We have also disclosed some existing open problem statements, ongoing research challenges that have relevant. This article should be beneficial to increase awareness of researchers and practitioners working toward Streaming Data Architectures and help understand existing challenges and advancements in the field of Real-Time Analytics.

References

1. Davoudian A, Liu M (2020) Big data systems: a software engineering perspective. ACM Comput Surv 53(5, Article 110):39. https://doi.org/10.1145/3408314
2. Dubuc T, Stahl F, Roesch EB (2021) Mapping the big data landscape: technologies, platforms and paradigms for real-time analytics of data streams. IEEE Access 9:15351–15374. https://doi.org/10.1109/ACCESS.2020.3046132
3. Al-Ghossein M, Abdessalem T, Barre A (2021) A survey on stream-based recommender systems. ACM Comput Surv 54(5, Article 104):36. https://doi.org/10.1145/3453443
4. Michel O, Bifulco R, Rétvári G, Schmid S (2021) The programmable data plane: abstractions, architectures, algorithms, and applications. ACM Comput Surv 54(4, Article 82):36. https://doi.org/10.1145/3447868
5. Ellis B (2014) Real-time analytics: techniques to analyze and visualize streaming data. Wiley & Sons
6. Peng J, Cai K, Jin X (2020) High concurrency massive data collection algorithm for IoMT applications, Comput Commun 157:402–409. https://doi.org/10.1016/j.comcom.2020.04.045. ISSN 0140-3664
7. Chen G, Yeh T, Liu C, İk T (2020) Microscopic traffic monitoring and data collection cloud platform based on aerial video. In: 2020 IEEE wireless communications and networking conference (WCNC), pp 1–6. https://doi.org/10.1109/WCNC45663.2020.9120623
8. Nougnanke KB, Labit Y (2020) Novel adaptive data collection based on a confidence index in SDN. In: 2020 IEEE 17th annual consumer communications & networking conference (CCNC), pp 1–6. https://doi.org/10.1109/CCNC46108.2020.9045207
9. Melenli S, Topkaya A (2020) Real-time maintaining of social distance in Covid-19 environment using image processing and Big Data. In: 2020 Innovations in intelligent systems and applications conference (ASYU), pp 1–5. https://doi.org/10.1109/ASYU50717.2020.9259891
10. Facebook. [n.d.]. Facebook fbthrift. Retrieved from https://github.com/facebook/fbthrift/
11. Google. [n.d.]. Protocol Buffers. Retrieved from https://developers.google.com/protocol-buffers/
12. Apache. [n.d.]. Welcome to Apache Avro!. Retrieved from https://avro.apache.org/
13. JSON. [n.d.]. Introducing JSON. Retrieved from https://www.json.org/
14. Silvestre PF, Fragkoulis M, Spinellis D, Katsifodimos A (2021) Clonos: consistent causal recovery for highly-available streaming dataflows. In: Proceedings of the 2021 international

conference on management of data. Association for Computing Machinery, New York, NY, USA, pp 1637–1650. https://doi.org/10.1145/3448016.3457320
15. Ma Q, Gu Y, Lee W-C, Yu G, Liu H, Wu X (2020) REMIAN: real-time and error-tolerant missing value imputation. ACM Trans Knowl Discov Data 14(6, Article 77):38. https://doi.org/10.1145/3412364
16. Lesniak A, Laigner R, Zhou Y (2021) Enforcing consistency in microservice architectures through event-based constraints. In: Proceedings of the 15th ACM international conference on distributed and event-based systems. Association for Computing Machinery, New York, NY, USA, pp 180–183. https://doi.org/10.1145/3465480.3467839
17. Apache. [n.d.]. Apache Samza. Retrieved from https://projects.apache.org/project.html?samza
18. Apache. [n.d.]. Apache Helix. Retrieved from https://projects.apache.org/project.html?helix
19. Apache. [n.d.]. Apache Kafka. Retrieved from https://projects.apache.org/project.html?kafka
20. Apache. [n.d.]. Apache Flume. Retrieved from https://projects.apache.org/project.html?flume
21. Wang G, Chen L, Dikshit A, Gustafson J, Chen B, Sax MJ, Roesler J, Blee-Goldman S, Cadonna B, Mehta A, Madan V, Rao J (2021) Consistency and completeness: rethinking distributed stream processing in Apache Kafka. In: Proceedings of the 2021 international conference on management of data (SIGMOD/PODS '21). Association for Computing Machinery, New York, NY, USA, pp 2602–2613. https://doi.org/10.1145/3448016.3457556
22. Lv Z, Chen D, Singh AK (2021) Big data processing on volunteer computing. ACM Trans Internet Technol 21(4, Article 83):20. https://doi.org/10.1145/3409801
23. Lin C, Ouyang Z, Wang X, Li H, Huang Z (2021) Preserve integrity in realtime event summarization. ACM Trans Knowl Discov Data 15(3, Article 49):29. https://doi.org/10.1145/3442344
24. Wu H, Ma T, Wu L, Xu F, Ji (2021) Exploiting heterogeneous graph neural networks with latent worker/task correlation information for label aggregation in crowdsourcing. ACM Trans Knowl Discov Data 16(2, Article 27):18. https://doi.org/10.1145/3460865
25. Rozet A, Poppe O, Lei C, Rundensteiner EA (2020) Muse: multi-query event trend aggregation. In: Proceedings of the 29th ACM international conference on information & knowledge management. Association for Computing Machinery, New York, NY, USA, pp 2193–2196. https://doi.org/10.1145/3340531.3412138
26. Traub J, Grulich PM, Cuéllar AR, Breß S, Katsifodimos A, Rabl T, Markl V (2021) Scotty: general and efficient open-source window aggregation for stream processing systems. ACM Trans Database Syst 46(1, Article 1):46.https://doi.org/10.1145/3433675
27. Apache. [n.d.]. Apache Storm. Retrieved from https://projects.apache.org/project.html?storm
28. Apache. [n.d.]. Apache Flink. Retrieved from https://projects.apache.org/project.html?flink
29. Apache. [n.d.]. Apache Spark. Retrieved from https://projects.apache.org/project.html?spark
30. Apache. [n.d.]. Apache Fluo. Retrieved from https://projects.apache.org/project.html?fluo
31. Apache. [n.d.]. Apache Tez. Retrieved from https://projects.apache.org/project.html?tez
32. Körber M, Glombiewski N, Seeger B (2021) Index-accelerated pattern matching in event stores. In: Proceedings of the 2021 international conference on management of data. Association for Computing Machinery, New York, NY, USA, pp 1023–1036. https://doi.org/10.1145/3448016.3457245
33. Zheng Q, Cranor CD, Jain A, Ganger GR, Gibson GA, Amvrosiadis G, Settlemyer BW, Grider G (2020) Streaming data reorganization at scale with DeltaFS indexed massive directories. ACM Trans Storage 16(4, Article 23):31. https://doi.org/10.1145/3415581
34. Izadpanah R, Peterson C, Solihin Y, Dechev D (2021) PETRA: persistent transactional non-blocking linked data structures. ACM Trans Archit Code Optim 18(2, Article 23):26. https://doi.org/10.1145/3446391
35. Srikanth S, Jain A, Conte TM, Debenedictis EP, Cook J (2021) SortCache: intelligent cache management for accelerating sparse data workloads. ACM Trans Archit Code Optim 18(4, Article 56):24. https://doi.org/10.1145/3473332
36. Singh A, Dave S, Zardoshti P, Brotzman R, Zhang C, Guo X, Shrivastava A, Tan G, Spear M (2021) SPX64: a scratchpad memory for general-purpose microprocessors. ACM Trans Archit Code Optim 18(1, Article 14):26. https://doi.org/10.1145/3436730

37. Maass S, Kumar MK, Kim T, Krishna T, Bhattacharjee A (2020) ECO TLB: eventually consistent TLBs. ACM Trans Archit Code Optim 17(4, Article 27):24. https://doi.org/10.1145/3409454
38. Apache. [n.d.]. Apache CouchDB. Retrieved from https://projects.apache.org/project.html?couchdb
39. Apache. [n.d.]. Apache Carbondata. Retrieved from https://projects.apache.org/project.html?carbondata
40. Apache. [n.d.]. Apache Phoenix. Retrieved from https://projects.apache.org/project.html?phoenix
41. Apache. [n.d.]. Apache Lucene-core. Retrieved from https://projects.apache.org/project.html?lucene-core
42. Paschalides D, Stephanidis D, Andreou A, Orphanou K, Pallis G, Dikaiakos MD, Markatos E (2020) MANDOLA: a big-data processing and visualization platform for monitoring and detecting online hate speech. ACM Trans Internet Technol 20(2, Article 11):21. https://doi.org/10.1145/3371276
43. Farhat O, Daudjee K, Querzoni L (2021) Klink: progress-aware scheduling for streaming data systems. In: Proceedings of the 2021 international conference on management of data. Association for Computing Machinery, New York, NY, USA, pp 485–498. https://doi.org/10.1145/3448016.3452794
44. Li Y, Li K, Chen C, Zhou X, Zeng Z, Li K (2021) Modeling temporal patterns with dilated convolutions for time-series forecasting. ACM Trans Knowl Discov Data 16(1, Article 14):22. https://doi.org/10.1145/3453724
45. Savva F, Anagnostopoulos C, Triantafillou P, Kolomvatsos K (2020) Large-scale data exploration using explanatory regression functions. ACM Trans Knowl Discov Data 14(6, Article 76):33. https://doi.org/10.1145/3410448
46. Apache. [n.d.]. Apache Zeppelin. Retrieved from https://projects.apache.org/project.html?zeppelin
47. Grafana Labs [n.d.]. Grafana Labs. Retrieved from https://grafana.com/
48. Prometheus [n.d.]. Prometheus. Retrieved from https://prometheus.io/
49. Zang T, Zhu Y, Xu Y, Yu J (2021) Jointly modeling spatio–temporal dependencies and daily flow correlations for crowd flow prediction. ACM Trans Knowl Discov Data 15(4, Article 58):20. https://doi.org/10.1145/3439346
50. Deng J, Chen X, Fan Z, Jiang R, Song X, Tsang IW (2021) The pulse of urban transport: exploring the co-evolving pattern for spatio-temporal forecasting. ACM Trans Knowl Discov Data 15(6, Article 103):25. https://doi.org/10.1145/3450528
51. Daghistani A, Aref WG, Ghafoor A, Mahmood AR (2021) SWARM: adaptive load balancing in distributed streaming systems for big spatial data. ACM Trans Spatial Algorithms Syst 7(3, Article 14):43. https://doi.org/10.1145/3460013
52. Fu Y, Soman C (2021) Real-time data infrastructure at Uber. In: Proceedings of the 2021 international conference on management of data. Association for Computing Machinery, New York, NY, USA, pp 2503–2516. https://doi.org/10.1145/3448016.3457552
53. Badr E (2021) Images in space and time: real big data in healthcare. ACM Comput Surv 54(6, Article 113):38. https://doi.org/10.1145/3453657
54. Can YS, Ersoy C (2021) Privacy-preserving federated deep learning for wearable IoT-based biomedical monitoring. ACM Trans Internet Technol 21(1, Article 21):17. https://doi.org/10.1145/3428152
55. Yue Z, Ding S, Zhao L, Zhang Y, Cao Z, Tanveer M, Jolfaei A, Zheng X (2021) Privacy-preserving time-series medical images analysis using a hybrid deep learning framework. ACM Trans Internet Technol 21(3, Article 57):21. https://doi.org/10.1145/3383779
56. Piccialli F, Giampaolo F, Prezioso E, Crisci D, Cuomo S (2021) Predictive analytics for smart parking: a deep learning approach in forecasting of IoT data. ACM Trans Internet Technol 21(3, Article 68):21. https://doi.org/10.1145/3412842
57. Djenouri Y, Djenouri D, Lin JC-W (2021) Trajectory outlier detection: new problems and solutions for smart cities. ACM Trans Knowl Discov Data 15(2, Article 20):28. https://doi.org/10.1145/3425867

58. Shahid H, Shah MA, Almogren A, Khattak HA, Din IU, Kumar N, Maple C (2021) Machine learning-based mist computing enabled internet of battlefield things. ACM Trans Internet Technol 21(4, Article 101):26. https://doi.org/10.1145/3418204
59. Shen J, Cao J, Lederman O, Tang S, Pentland AS (2021) User profiling based on nonlinguistic audio data. ACM Trans Inf Syst 40(1, Article 17):23. https://doi.org/10.1145/3474826
60. Almaslukh A, Kang Y, Magdy A (2021) Temporal geo-social personalized keyword search over streaming data. ACM Trans Spatial Algorithms Syst 7(4, Article 20): 28. https://doi.org/10.1145/3473006
61. Hidalgo JIG, Santos SGTC, Barros RSM (2021) Dynamically adjusting diversity in ensembles for the classification of data streams with concept drift. ACM Trans Knowl Discov Data 16(2, Article 31):20. https://doi.org/10.1145/3466616
62. Wu H, Wu Q, Ng MK (2021) Knowledge preserving and distribution alignment for heterogeneous domain adaptation. ACM Trans Inf Syst 40(1, Article 16):29. https://doi.org/10.1145/3469856
63. Kallas K, Niksic F, Stanford C, Alur R (2020) DiffStream: differential output testing for stream processing programs. Proc ACM Program Lang 4(OOPSLA, Article 153):29. https://doi.org/10.1145/3428221

Digital Video Steganography: An Overview

Suganthi Kumar and Rajkumar Soundrapandiyan

Abstract The recent evolution of digital communication demands securing the integrity of data transfer. Under such necessity, the data hiding practice, steganography allures the attention of security and forensic departments. Steganography is broadly classified into two types depending on the linguistic and technical features. Based on the domain in which the steganography is applied, the technical steganography is categorized into the following types: digital media, network, hardware, circuitry, and genome. Concerning the cover medium, digital media steganography is further classified into four types: image, audio, video, and documents. By widespread usage of video-based applications, video-specific techniques have recently got more emphasis. This rapid growth motivated the authors to provide a comprehensive literature review of this subject. This paper mainly talks about video steganography techniques in the recent decade. In addition, quality metrics used in the performance evaluation process are also discussed. This review is mainly focused on providing an overall guide for new researchers approaching the field of steganography.

Keywords Digital communication · Steganography · Secret communication · Technical steganography · Digital media · Video steganography · Quality metrics

1 Introduction

Communication is an essential part of human life. It was done in different ways at different times based on civilization. However, distance communication had always been challenging in ancient times. But in the present modern age with the advent of communication technologies, remote communication turned out to be much easier.

S. Kumar (✉) · R. Soundrapandiyan
School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu 632014, India
e-mail: k.suganthi@vit.ac.in

R. Soundrapandiyan
e-mail: rajkumarsrajkumar@gmail.com; rajkumars@vit.ac.in

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7_42

Wherein, digital media communication has become inevitable transferring multimedia files and messages between sender and receiver. Howbeit, excessive usage and over-reliance on the same pose several security risks and adversaries such as data intrusion, modification, theft, and impersonation, by unethical third parties, and it leads to having disastrous consequences for the information system. Thus, the process of securing the communication is as important as the communication itself.

There are many security mechanisms available for secure communication. Cryptography is one such approach that converts the message in such a way that it becomes ambiguous to the third party. Yet, it is not sufficient these days as the meaningless content itself draws attention. In such a scenario, steganography, the art and science of secret communication, is the optimal solution. Steganography mechanism establishes secret communication by hiding secret messages into some cover data. Thus, it becomes impossible for the eavesdropper to discern secret communication as steganography conceals its very existence. Nowadays, steganography is attainable for any end users with Internet access. At times, it is preferable to combine cryptography techniques with the steganography system to provide an additional layer of security. But using these techniques separately is not ideal. Watermarking is another data embedding mechanism that is proximate to steganography, which nevertheless has different goals and attributes, i.e., the existence of secret communication is often visible to everyone under certain conditions. The main property of watermarking is to prevent the ownership infringement of digital content. Thus, steganography strives for imperceptible communication to the human visual system (HVS), whereas watermarking puts robustness as a top priority.

The word steganography is the English rendition of the Greek word 'steganographia'. It means covered (steganos) writing (-graphia). As the name implies, steganography hides the secret messages in some cover media. Thus, it becomes impossible for the eavesdropper to discern secret communication as steganography conceals its very existence. Figure 1 portrays the working mechanism of the steganography. In the practice of steganography, the steganographers such as sender and receiver exchange secret messages using a cover medium through a network. The sender employs some data hiding process to replace the cover data with secret data. At this stage, a 'stego-key' can be used in encrypting the secret data before embedding it. Once the encrypted secret data is hidden into the cover medium, it is known as stego-medium, and it is sent to the receiver end via a network. The authorized receiver recovers the secret data from the stego-medium using a data retrieval process. The data retrieval algorithm is often the reversed version of the data hiding process. At times, some third-party steganalyst observes the stego-medium exchange between the steganographers and tries to remove or take out the secret data without the knowledge of steganographers. This action by steganalyst is called steganalysis. Steganalysis is the contrary mechanism to steganography: the art and science of detecting the hidden communication in a cover medium [1]. Steganalysis can be performed in two ways such as active steganalysis and passive steganalysis. The former procedure finds out only the presence (flag-1) or absence (flag-0) of the hidden message. Whereas, the latter tries to extract the entire secret data or certain characteristics of its length. Therefore, the primary motive of the steganography

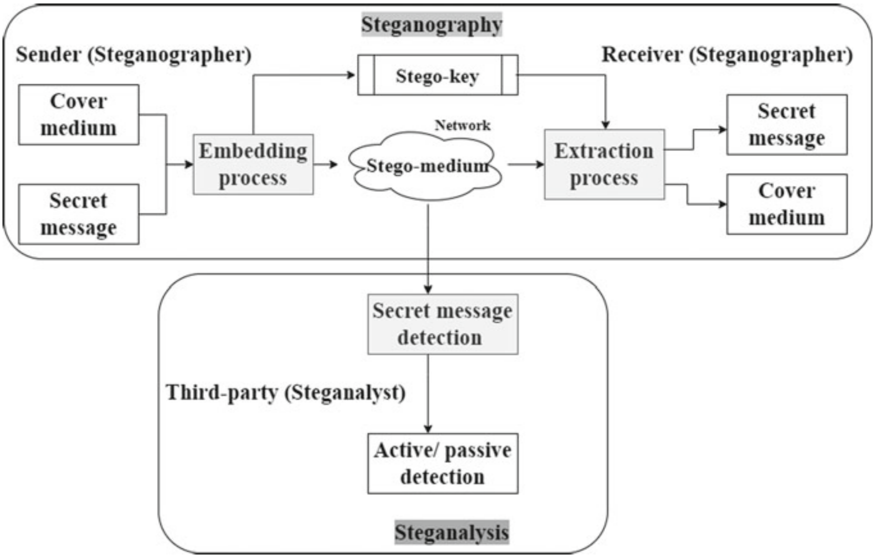


Fig. 1 Representation of steganography and steganalysis mechanisms

algorithm is to devise a secret communication framework that is not detectable by any steganalysis tools. Though the steganography technique conceals the very existence of the secret message from HVS, attacks on them are still possible. Because no matter what the embedding process is, it slightly alters certain properties of the cover medium. This causes quality attenuation in the final stego-medium which may not be noticeable by the HVS but can easily give a clue to some strong steganalysis mechanisms. Hence, the quality of the stego-medium needs to be standardized using computational evaluation metrics. In steganography, the performance or quality is generally measured in terms of three aspects such as imperceptibility, hiding capacity, and robustness.

Imperceptibility: The visual quality of the stego-medium. The cover medium and stego-medium should be identical.

Hiding capacity: The accommodation of secret data in cover medium without visual degradation.

Robustness: The resistance of the secret data inside the cover medium against any kind of attacks. And the secret data should be recovered by the receiver without data attenuation.

However, there are some more prerequisites such as reversibility and computational complexity are also required for better performance. Reversibility is the lossless or absolute retrieval of the secret data at the receiver side. And, computational complexity is the cost of time and space taken for the successful execution of the steganography algorithm which needs to be lesser for better performance. There are many quality metrics and tools available to measure the performance of steganographic systems. Besides, it is necessary to install well-qualified steganographic systems where the topmost secret communication is indispensable. Medical, military and law enforcement, geographical, entertainment, government, corporate, intelligence agencies, and banking are some of the domains where steganography systems can be applied.

Steganography has become a promising field of research in the latest decade. This rapid growth persuaded the authors to provide a comprehensive review of steganography. This paper mainly presents some recent trends and techniques utilized in the steganography domain. In addition, quality metrics and tools used in the performance evaluation process are also discussed in this paper.

The remaining sections of the paper are as follows: Sect. 2 presents an overview of steganography methods. Section 3 discusses the performance measures, and Sect. 4 concludes the paper.

2 Steganography

The importance of secure communication is raising lately due to data breaches on the Internet and social networks. Steganography is the optimal choice for securing and storing secret information/data. An ideal and perfect steganography system must attain the prerequisites of steganography which include imperceptibility, capacity, and robustness. In addition to these, there are some more requirements such as reversibility and computational complexity for better performance. As mentioned earlier, steganography is the practice of hiding secret data within a cover medium. Based upon the type of cover medium and area in which steganography is deployed, it is broadly categorized into two types as linguistic steganography and technical steganography.

Linguistic steganography is text steganography where written communication language features are utilized to hide secret messages in some non-obvious ways [2]. It is further classified as semantic based and syntax based. The former uses synonymous words or phrases for concealing the secret messages [3, 4]. However, the latter exploits punctuating elements such as full stop (.), comma (,), semicolon (;), and colon (:) for replacing the secret message [5]. Here, the embedding process is performed systematically without affecting the sentence meaning [6].

Next, the ***technical steganography*** deploys some scientific mechanisms to embed secret messages strategically. This steganography is further categorized into five main types such as network, hardware, circuitry, genome, and digital media.

Network steganography is the latest inclusion of the steganography family. The network steganography is further classified into two types as inter-protocol steganography and intra-protocol steganography [7]. Here, secret data is concealed inside redundant communication mechanisms, and also it creates an embedded communication channel using protocols like UDP, TCP, IP, etc.

A **hardware** type of steganography allows the secret data to be hidden into the space that is unused or holds residual data of a hard disk [8, 9]. Similarly, the software can also use the steganography approach by hiding secret information inside the code layout. In **circuitry** steganography, the secret data is concealed into an electronic circuit layout [10].

Genome steganography is an emerging technology that embeds the secret data behind the simulated biological structures of deoxyribonucleic acid (DNA). The high randomness in a DNA sequence is used for hiding purposes, where the encrypted secret data is placed as microdots between two primers [11, 12].

The standard and generally used type of steganography is **digital media** steganography. It embeds the secret message within multimedia files like images, video, audio, documents, etc.

Figure 2 exhibits the taxonomy and techniques of steganography. As the realm of steganography is vast as represented in Fig. 2, this paper presents the comprehensive literature only on digital video steganography techniques. Moreover, the recent year has been very progressive for video communication. Wherein, videos gain a wide spectrum of practical significance as they are being shared on social media frequently. There are scores of researches performed on video steganography using a variety of techniques. The following subsection explains video steganography techniques in detail.

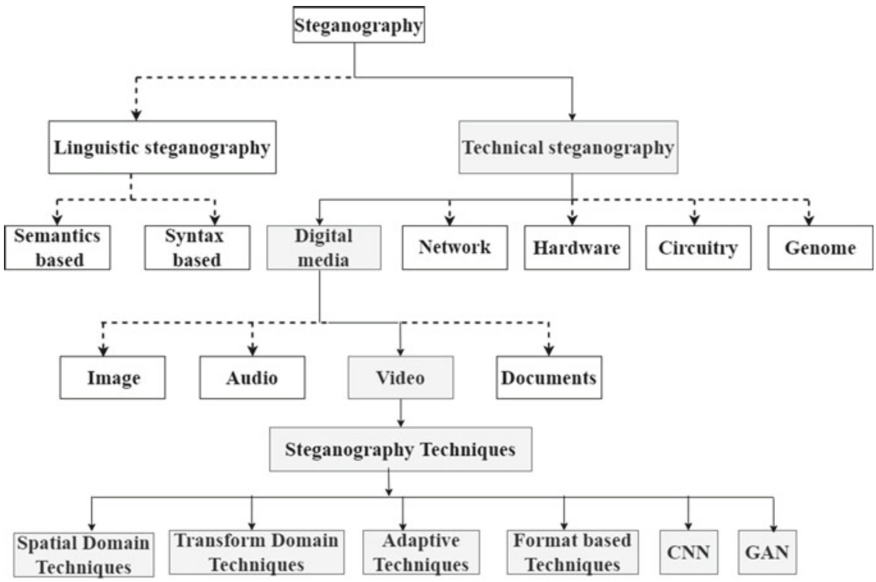


Fig. 2 Taxonomy and techniques of steganography

2.1 Digital Video Steganography

After carefully analyzing all the existing frameworks, the video steganography techniques are primarily classified into seven methods as (1) substitution techniques, (2) transform domain techniques, (3) adaptive techniques, (4) format-based techniques, (5) cover generation, (6) convolutional neural network (CNN), and (7) general adversarial network (GAN)-based methods.

Video steganography techniques. In video steganography, video content is utilized as a cover medium to embed the secret data. Video steganography has recently drawn more attention not only because of the security requirements, but also videos are more favored for the following reasons: (1) The high dimensionality in videos provides large space for data hiding, (2) the shot transitions in videos can be used for embedding purposes as they obscure to the HVS, and (3) the tampers in videos can be managed easily. (4) As videos are the sequence of frames/images, techniques that are feasible for image steganography can also be combined. There has been a lot of research performed on video steganography. The following subsections elaborate on the methods of video steganography and some related existing works of the reviewed techniques.

Substitution techniques. It is a spatial domain-based technique. Substitution is deployed in the color components RGB and YUV. Substitution technique includes the following: least significant bit (LSB), bit plane complexity segmentation (BPCS), and tri-way pixel value differencing (TPVD), etc. The LSB technique is one of the widely used substitution methods. This method displaces certain LSBs of the cover data by the bits of secret data. For instance, Chen and Qu [13] developed a video steganography framework embedding the secret bits into the LSBs of pixels from RGB components. Many video steganography-based research works have also employed encryption of secret data before the hiding process [14, 15]. Despite being simple and having a higher payload capacity, LSB method sometimes deteriorates the quality as more significant bits are utilized. This disadvantage was overcome by the evolution methods BPCS and TPVD. BPCS method is benefitted from the fact that the HVS is imperceptible in some color regions. Thus, the BPCS decomposes the frame/image of the video into bit planes. After decomposing, based on the complexity of the regions, the secret data is hidden in minimum perceived quality regions. For example, Kumar et al. [16] developed an LSB substitution-based video steganography framework where the secret data is hidden into the bit plane 4-4-0 LSBs of R-G-B components of the frame. In this method, it is considered that the change in the blue component is more sensitive to HVS. Thus, only red and green components are utilized for embedding. The TPVD is another substitution-based technique. It is the advancement of the pixel value differencing (PVD) method. The PVD first finds the difference between two adjacent pixels and hides the secret data. This method has restricted to ranges; each range has length breadth and height. The TPVD method, in contrast, utilizes all the horizontal, vertical, and diagonal edges for embedding. This improves the capacity.

Transform domain techniques. Generally, transform domain methods are resilient against attacks but computationally intense. Despite its complexity, this method maintains the quality of the stego-medium. In this technique, firstly, each cover frame of the video is converted into its transform domain, then the secret data is concealed into the selected transformed coefficients. Next, the altered transform coefficients are inversed to form the original cover frame format. The transform domain methods include discrete Fourier transform (DFT), discrete cosine transform (DCT), and discrete wavelet transform (DWT). Lately, some transform methods like complex wavelet transform (CWT), integer wavelet transform (IWT), and dual-tree complex wavelet transform are also utilized by the researchers. Here, the DFT method is not so used in steganography as it tends to generate round-off errors [17]. However, the DCT method is very famous in video steganography and widely used in video compression techniques. The main notion of this technique is to hide the secret data behind the compressed domain in a block-wise or frame-wise manner. MPEG-1, MPEG-2, and H.263 are some of the DCT-based compression techniques. Recently, the format H.264/AVC is utilized for the video steganography process. Earlier, Chae et al. [18] presented a data hiding process in MPEG-2 videos. This method initially transforms the secret data and cover content to the DCT domain. Then the quantized coefficients of the secret message are encoded using multidimensional lattice and embedded into the cover DCT coefficients. Similarly, Yang et al. [19] worked on H. 246/AVC compressed videos, where the cover video is converted into a YUV color component. Then the secret data is hidden bit by bit in the DCT domain of the color component. DWT has achieved much attention in the fields of signal processing and video steganography. The decomposed wavelets provide a multi-resolution description and temporal resolution, where frequency and spatial information can be captured. In addition, as DWT does not segment the frame into non-overlapping blocks, it avoids the blocking artifacts. Ramalingam et al. [20] developed a change detection-based video steganography technique. They used DCT coefficients to determine the scene change in the frames and the DWT domain to embed the secret. Hence, this method improves security and minimizes visual degradation.

Adaptive techniques. The adaptive technique targets specific regions or pixels for the embedding process. Combining adaptive methods and spatial and transform domains is an exceptional case [21–25]. Wherein, before the embedding process, the statistical characteristics of cover frames are observed. Based on these observed features, the suitable regions in the cover medium are adopted for the data hiding process. These adopted regions are called the Regions of Interest (ROI). Then, the secret data is hidden within spatial or transform or wavelet domains of the ROI. However, in the case of recognizing the ROI, the video steganography system gets benefitted from the fact that the HVS is deficient in discerning the changes in transient regions [15, 26, 27]. Considering this limitation, many research works have been presented. Hashemzadeh [28] used the motion clue feature points for their video steganography framework wherein the momentum of a particular key point is detected and tracked. And the spatial and temporal behaviors of the key point are observed and used for identifying the appropriate regions to hide the secret data. The secret data is

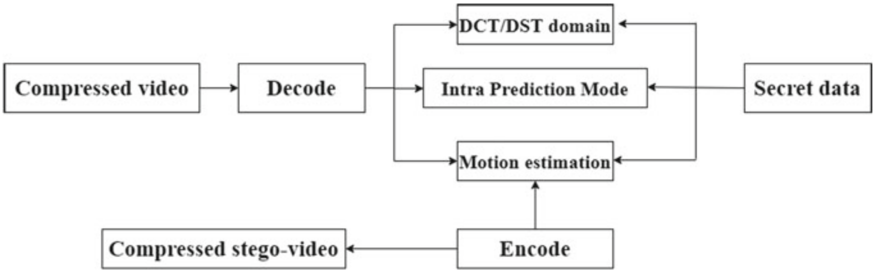


Fig. 3 Format-based steganography

embedded using the LSB substitution method. Similarly, Kumar et al. [29] proposed an optimal video steganography framework. This method determines dynamic key points using motion vectors and dilates them to form dynamic ROI, where the secret data is embedded. Utilizing key points of the frame region for the data hiding process is another fashion of the adaptive technique. Recently, Kumar et al. [16] deployed scale-invariant feature transform (SIFT) and speeded-up robust features (SURF) key points detection algorithms to find the ROI, where the encrypted secret data is embedded. Though this method is highly resilient against attacks, it performs poorly for uncompressed and low visual quality cover videos.

Format-based techniques. This is a video format-specific technique that uses the video coding standards for the data embedment process. Figure 3 is the general representation of the format-based steganography process. In this technique, the video coding processes such as intra-prediction, motion estimation, and DCT/DST coefficients are decoded, and the secret data is embedded in any of these processes. Once the secret data is embedded, it is again encoded to form the compressed version of the stego-video. The recent compression standard H.264/AVC gives a high video compression efficiency. This encoder is the most suited for network communication [30]. Many steganography techniques are designed for this video format. Flash video files (.FLV) are a simple and the most famous compression technique on social networks and have a better compression rate than other techniques. Mozo et al. [31] presented the uncomplicated FLV format-based steganographic system. This method uses the entire video tags of the file to embed the secret data. This addition does not modify the video/audio tags. Hence, the video quality is not affected and also provides huge embedding space. However, the imbalance between the cover and secret data size reduces tamper resistance.

Cover generation. All the above-mentioned steganography techniques use some carrier/cover medium to hide the secret message. However, in the cover generation techniques, the algorithm itself synthesizes a cover medium for communication. Sampat et al. [32] initiated this idea of generating dynamic cover videos. This process generates the cover video using a secret key and secret data itself. This process uses the function F(C, S), where F is the function to develop the cover file and C is the measure of sample required to hide the secret data S. This method acquires an image database to get frames for the cover video generation. However, this method is passive as the selected frame/image sequence of the cover video is unrelated to each other. This may give a clue to the third party. This may be overcome by having images that are slightly relevant to each other or the slideshow/animation representation of images can be used rather than having an inappropriate video and audio file combination.

Convolutional neural network (CNN). The CNN-based methods have become a trend and are widely used in many video steganographic systems. Because CNN can effectively be applied in various procedures such as image classifications/ recognition, language processing, medical image processing, and recommendation system. This method is completely benefitted from encoder–decoder architecture. Wherein, the secret data and cover video/frames are given as inputs to the encoder to form stego-video. The decoder extracts the embedded secret data. This is the fundamental idea in all the methods. However, the way secret data and cover data are integrated differs from one another. This architectural variation is made in the number of layers used, pooling layer, strides, the convolutional layer, kernel size, activation function utilized, and loss function, etc. The only concern here is the size of secret data, and cover data has to be the same. For example, U-Net architecture-based encoder and decoder model is employed in many works [33–35] for data hiding. The main advantage of U-Net architecture is the improved resolution of the stego-medium as upsampling operators are used. Van et al. [36] suggested a hiding technique using U-Net (H-Net) and a retrieval method using R-Net. Here, rectified linear unit (ReLU) and batch normalization are used. Similarly, Wu et al. [36] proposed a separable convolution with residual block (SCR) for the embedding purpose. This method also uses batch normalization and Exponential Linear Unit (ELU). Generally, in video steganography, 3D convolutional layers are used. Mishra et al. [37] presented VStegNET-based video steganography where the cover video and secret data is given to the autoencoder network to generate stego-video. The reversed network framework is deployed in extracting the secret data.

General adversarial network (GAN). This is the deep level of CNN [38]. The game theory (GT) is the main component in a GAN architecture for training the generative model utilizing the adversarial processes. In the steganographic process, the GAN architecture contains three models such as generator (G), discriminator (D), and steganalyzer (S). The functionalities of these components include generating stego-medium from the cover frame and secret data, comparing the cover frame and stego-medium, and checking the secret data is robust, respectively. In GAN architecture, the models G, D, and S are considered as players and subjected to compete against each other [29] to produce high-quality stego-medium. Here, player G gets updated from the errors of players D and S. To produce a realistic stego-medium that is closely similar to the cover frame, the error of D and S should be minimized between (0, 1). Volkhonsiy et al. [39] used DCGAN-based steganographic system. It is a simple architecture having three models—G, D, and D. Similar to this, Shi et al. [40] proposed GAN model that has four fractional convolutional layers and one functional layer that uses hyperbolic tangent activation with base (WGAN). In this method, all three players compete against each other until 'G' generates high-quality stego-medium, while player 'D' recovers the secret data from the stego-medium. However, player 'S' steganlyzes on the player 'G' to get the probability of hidden

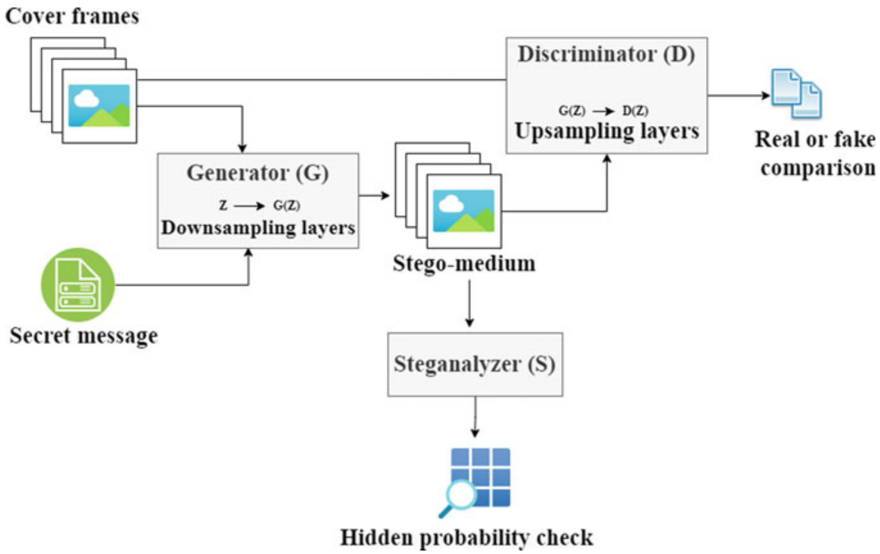


Fig. 4 Working model of GAN

information. Figure 4 is the general working algorithm of GAN as in [41]. Here, the generator fuses the frames of the cover video and secret message to produce stego-medium by downsampling the frames. The discriminator confronts the generator to find out the produced stego-medium is fake or not. The steganalyzer checks for the possibility of hidden information.

3 Quality Evaluation Techniques

In general, the performance of any steganography framework is examined in terms of imperceptibility, hiding capacity, and robustness. As the important requirement of steganography, imperceptibility must be analyzed meticulously. There are various metrics used for the evaluation of visual quality. The visual quality measurements are broadly classified into two types as follows: statistical error-based measurements and structural error-based measurements. The statistical error-based metrics calculate the quality value that differs from its standardized value. The difference between the calculated value and standard value is the error. Here, the error signal and the optical perception are reversely proportional to each other. These metrics use the mean square error to find the quality of the video. Mean square error (MSE), root mean square error, signal-to-noise ratio (SNR), peak signal-to-noise ratio (PSNR), weighted peak signal-to-noise ratio (WPSNR), video quality metric (VQM), normalized absolute error (NAE), normalized least square error (NLSE), and bit error rate (BER) are some of the statistical error-based metrics. Another category is structural error-based metrics. This method complies with the fact that the HVS can easily draw out structural information than statistical information. Structural similarity index measure (SSIM), structural content (SC), normalized cross-correlation (NCC), and zero normalized cross-correlation (ZNCC) are some of the examples of this kind. All the above-discussed methods find the values for each frame of the video and take the average of it. Moreover, these metrics were designed for images initially. Thus, some video-specific metrics such as video quality measurement (VQM), weighted SSIM ($SSIM^{wt}$), video quality model—general (VQM_g), and moving pictures quality metric (MPQM) are designed to check particularly the video quality.

Robustness is another important requirement in steganographic methods. Wherein, robustness means the ability to recover the secret message at the receiver end without having any data loss despite the attacks in the network. Hence in the robustness analysis process, attacks are imposed on stego-video, and then the secret data is extracted to test its robustness. The metrics employed for imperceptibility analysis can also be used in robustness analysis too. The only difference is the input medium. In imperceptibility analysis, the entire cover medium and stego-medium are given as input data for all the metrics. Herein, robustness analysis, the original secret data, and the extracted secret data are the inputs for the metrics.

Next, the hiding capacity is yet another important feature that should be analyzed. Capacity is defined as the rate of secret data that can be embedded within the cover medium. Thus, hiding capacity calculates the secret data distribution rate across the cover pixels. Hence, the unit of hiding rate is bit per pixel (BPP). This calculation can be performed in two ways: overall capacity (HR_{oc}) and algorithm-provided capacity (HR_{ac}). The overall capacity is the hiding ratio of the secret data to the entire cover medium. On the other hand, the algorithm-provided capacity estimates the hiding rate within the Region of Interest (ROI), where the pixels are chosen for the embedding process by the steganography algorithm.

4 Conclusion

This paper presented an overview of steganography and its types and techniques. The paper was introduced with the significance, origin, description, and application of steganography. Then, the broad classifications of steganography based on the cover type were discussed. Various methods of steganography were presented, and technicality-based steganography methods were given special attention. Technical steganography is categorized into five major types: digital media, network, hardware, circuitry, and genome. This paper adopted video-specific digital media steganography techniques for reviewing. After carefully observing all the existing video steganography frameworks, the techniques are majorly grouped into seven types such as (1) substitution techniques, (2) transform domain techniques, (3) adaptive techniques, (4) format-based techniques, (5) cover generation, (6) convolutional neural network (CNN), and (7) general adversarial network (GAN). Each technique was thoroughly scrutinized and precisely presented in the paper. Finally, the quality evaluation techniques used in analyzing the steganography framework were discussed.

From this survey, we noticed that all the techniques have their pros and cons depending upon the compatibility between the algorithm and the application in which it gets applied. Moreover, none of the reviewed techniques attains all the mentioned prerequisites such as imperceptibility, hiding capacity, and robustness. There is always some sort of imbalance between them. So it is recommended for the new steganography aspirant to focus more on providing a trade-off between these quality requirements while maintaining the algorithm and application agreement.

References

1. Patel A, Shah M, Chandramouli R, Subbalakshmi KP (2007) Covert channel forensics on the internet: issues, approaches, and experiences. Int J Netw Secur 5(1):41–50
2. Hamzah AA, Khattab S, Bayomi H (2019) A linguistic steganography framework using Arabic calligraphy. J King Saud Univ Comput Inf Sci
3. Shirali-Shahreza MH, Shirali-Shahreza M (2008) A new synonym text steganography. In: 2008 International conference on intelligent information hiding and multimedia signal processing. IEEE, pp 1524–1526
4. Wang F, Huang L, Chen Z, Yang W, Miao H (2013) A novel text steganography by context-based equivalent substitution. In: 2013 IEEE international conference on signal processing, communication and computing (ICSPCC 2013). IEEE, pp 1–6
5. Shirali-Shahreza M (2008) Text steganography by changing words spelling. In: 2008 10th international conference on advanced communication technology, vol 3. IEEE, pp 1912–1913
6. Murphy B, Vogel C (2007) The syntax of concealment: reliable methods for plain text information hiding. In: Security, steganography, and watermarking of multimedia contents IX, vol 6505. International Society for Optics and Photonics, p 65050Y
7. Lubacz J, Mazurczyk W, Szczypiorski K (2014) Principles and overview of network steganography. IEEE Commun Mag 52(5):225–229
8. Neuner S, Voyiatzis AG, Schmiedecker M, Brunthaler S, Katzenbeisser S, Weippl ER (2016) Time is on my side: Steganography in filesystem metadata. Digit Investig 18:S76–S86
9. Khan H, Javed M, Khayam SA, Mirza F (2011) Designing a cluster-based covert channel to evade disk investigation and forensics. Comput Secur 30(1):35–49
10. Johnson NF, Duric Z, Jajodia S (2001) Information hiding: steganography and watermarking-attacks and countermeasures: steganography and watermarking: attacks and countermeasures, vol 1. Springer Science & Business Media
11. Wang Z, Zhao X, Wang H, Cui G (2013) Information hiding based on DNA steganography. In: 2013 IEEE 4th international conference on software engineering and service science. IEEE, pp 946–949
12. Risca VI (2001) DNA-based steganography. Cryptologia 25(1):37–49
13. Chen S, Qu Z (2018) Novel quantum video steganography and authentication protocol with large payload. Int J Theor Phys 57(12):3689–3701
14. Sudeepa KB, Raju K, HS RK, Aithal G (2016) A new approach for video steganography based on randomization and parallelization. Procedia Comput Sci 78:483–490
15. Manisha S, Sharmila TS (2019) A two-level secure data hiding algorithm for video steganography. Multidimension Syst Signal Process 30(2):529–542
16. Kumar S, Soundrapandiyan R (2020) Robust approach of video steganography using combined keypoints detection algorithm against geometrical and signal processing attacks. J Electron Imaging 29(4):043007

17. Raja KB, Chowdary CR, Venugopal KR, Patnaik LM (2005) A secure image steganography using LSB, DCT and compression techniques on raw images. In: 2005 3rd international conference on intelligent sensing and information processing. IEEE, pp 170–176
18. Chae JJ, Manjunath BS (1999) Data hiding in video. In: Proceedings 1999 international conference on image processing (Cat. 99CH36348), vol 1. IEEE, pp 311–315
19. Yang M, Bourbakis N (2005) A high bitrate information hiding algorithm for digital video content under H. 264/AVC compression. In: 48th Midwest symposium on circuits and systems. IEEE, pp 935–938
20. Ramalingam M, Isa NAM (2016) A data-hiding technique using scene-change detection for video steganography. Comput Electr Eng 54:423–434
21. Mstafa RJ, Elleithy KM, Abdelfattah E (2017) A robust and secure video steganography method in DWT-DCT domains based on multiple object tracking and ECC. IEEE access 5:5354–5365
22. Singh D, Singh B (2015) Data hiding in videos using background subtraction. In: 2015 2nd International conference on recent advances in engineering & computational sciences (RAECS). IEEE, pp 1–5
23. Chung KL, Chiu CY, Yu TY, Huang PL (2017) Temporal and spatial correlation-based reversible data hiding for RGB CFA videos. Inf Sci 420:386–402
24. Mstafa RJ, Elleithy KM (2016) A video steganography algorithm based on Kanade-Lucas-Tomasi tracking algorithm and error correcting codes. Multimedia Tools Appl 75(17):10311–10333
25. Mansouri J, Khademi M (2009) An adaptive scheme for compressed video steganography using temporal and spatial features of the video signal. Int J Imaging Syst Technol 19(4):306–315
26. Fernando WAC (2006) Sudden scene change detection in compressed video using interpolated macroblocks in B-frames. Multimedia Tools Appl 28(3):301–320
27. Rajalakshmi K, Mahesh K (2018) Robust secure video steganography using reversible patch-wise code-based embedding. Multimedia Tools Appl 77(20):27427–27445
28. Hashemzadeh M (2018) Hiding information in videos using motion clues of feature points. Comput Electr Eng 68:14–25
29. Kumar S, Soundrapandiyan R (2021) A multi-image hiding technique in dilated video regions based on cooperative game-theoretic approach. J King Saud Univ Comput Inf Sci
30. Juurlink B, Alvarez-Mesa M, Chi CC, Azevedo A, Meenderinck C, Ramirez A (2012) Understanding the application: an overview of the h. 264 standard. In: Scalable parallel programming applied to H. 264/AVC decoding, pp 5–15
31. Mozo AJ, Obien ME, Rigor CJ, Rayel DF, Chua K, Tangonan G (2009) Video steganography using flash video (FLV). In: 2009 IEEE instrumentation and measurement technology conference. IEEE, pp 822–827
32. Sampat V, Dave K, Madia J, Toprani P (2012) A novel video steganography technique using dynamic cover generation. In: National conference on advancement of technologies–information systems & computer networks (ISCON–2012), Proceedings published in Int J Comput Appl (IJCA)
33. Wu P, Yang Y, Li X (2018) Image-into-image steganography using deep convolutional network. In: Pacific rim conference on multimedia. Springer, Cham, pp 792–802
34. Wu P, Yang Y, Li X (2018) Stegnet: mega image steganography capacity with deep convolutional network. Future Internet 10(6):54
35. Duan X, Jia K, Li B, Guo D, Zhang E, Qin C (2019) Reversible image steganography scheme based on a U-Net structure. IEEE Access 7:9314–9323
36. Van TP, Dinh TH, Thanh TM (2019) Simultaneous convolutional neural network for highly efficient image steganography. In: 2019 19th International symposium on communications and information technologies (ISCIT). IEEE, pp 410–415
37. Mishra A, Kumar S, Nigam A, Islam S (2019) VStegNET: video steganography network using spatio-temporal features and micro-bottleneck. In: BMVC, p 274
38. Pinson MH, Wolf S (2004) A new standardized method for objectively measuring video quality. IEEE Trans Broadcast 50(3):312–322

39. Volkhonskiy D, Borisenko B, Burnaev E (2016) Generative adversarial networks for image steganography
40. Shi H, Dong J, Wang W, Qian Y, Zhang X (2017) SSGAN: secure steganography based on generative adversarial networks. In: Pacific rim conference on multimedia. Springer, Cham, pp 534–544
41. Hayes J, Danezis G (2017) Generating steganographic images via adversarial training. arXiv preprint arXiv:1703.00371

Author Index

V. K. Asari et al. (eds.), *Computational Methods and Data Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 139,
https://doi.org/10.1007/978-981-19-3015-7