ISSN : 2582-7464

Computational Intelligence and Machine Learning

Volume No. - 5

Issue No. - 3

September - December 2024

ENRICHED PUBLICATIONS PVT.LTD

JE - 18,Gupta Colony, Khirki Extn,
Malviya Nagar, New Delhi - 110017.
E- Mail: info@enrichedpublication.com
Phone :- +91-8877340707

ISSN : 2582-7464

Computational Intelligence and Machine Learning

Aims and Scope

The primary objective of the Computational Intelligence and Machine Learning is to serve as a comprehensive, open-access platform that is dedicated solely to facilitating the progress and advancement of the field of Artificial Intelligence & Machine Learning by -

offering gifted and talented researchers engaged within the domain of Artificial Intelligence & Machine Learning a unique setting for them to get their work published and elevate their reputations/standing within the global community, as well as,

providing professionals, students, academics, and scholars free access to the latest and most advanced research outcomes, findings, and studies, being carried out in the field of Artificial Intelligence & Machine Learning, all across the world.

Related Topics

Soft computing

Fuzzy Logic

Artificial Neural Networks

Evolutionary Computing

Artificial Intelligence and Machine Learning

Artificial Immune Systems

Probabilistic Methods

Cognitive Robotics

Data mining

Computational Intelligence Methods for Bioinformatics and Biostatistics

ISSN : 2582-7464

Other emerging topics in Computational Intelligence

Nanobioscience

Information Forensics and technology

Nanotechnology

Cybersecurity

Big Data

Bioengineering and Biotechnology

Computational Neuroscience

ISSN : 2582-7464

Advisor

Editor-in-chief

Editorial Board Members

ISSN : 2582-7464

ISSN : 2582-7464

ISSN : 2582-7464

Computational Intelligence and Machine Learning

(Volume No. 5, Issue No. 3, Sep - Dec 2024)

Contents

Marathi Extractive Text Summarization using Latent Semantic Analysis and Fuzzy Algorithms

Virat V Giri1*, Dr. M.M. Math2, Dr. U.P. Kulkarni3
1 Principal, Sanjay Ghodawat Polytechnic, Kolhapur, India
2Professor, Dept of Comp Science and Engg, KLS, Gogte Institute of Technology, Belgaum, India
3 Professor, Dept. of Comp Science and Engg, SDM, College of Engineering and Technology, Dharwad, India

A B S T R A C T

Extractive text summarization involves the retention of only the most important sentences in a document. In the past, multiple approaches involving both statistical and machine learning-based methods have been used for this task. The crucial step in extractive text summarization is getting the right ranking order of sentences in the document in terms of their importance. Singular value decomposition or SVD algorithm based on latent semantic analysis focuses on recognizing the sections in the document which are related in terms of their semantic nature. Fuzzy algorithms involve reasoning of the priority order of the sentences using fuzzy logic unlike the use of discrete values. While significant work has been done for extractive text summarization in English and other foreign languages, there is ample scope for improving the performance of systems when dealing with Marathi text. In this paper, SVD and fuzzy algorithms are proposed for performing extractive text summarization on Marathi documents. Work is done upon the modeling principle, data flow, and parameters of these algorithms such that they are best suited for the task. An analysis of the characteristics of both these techniques is conducted to compare their benefits and shortcomings. The performance of both the algorithms is evaluated on a document dataset using standard performance metrics including the ROUGE metric. An unbiased comparison of both these techniques is carried out to inform the applicability of them, especially when working with Marathi or in general, non-English text.

Keywords Extractive Text Summarization, Singular Value Decomposition, Fuzzy Logic, Marathi Text.

INTRODUCTION

Natural language processing involves processing and modeling of natural language data to improve understanding of computers while ensuring that the semantic and syntactic structure of the data is retained. Text summarization is an important application of natural language processing which focuses on automatically deriving the summary of entered documents. There are two possible types of text summarization: abstractive text summarization and extractive text summarization.

Extractive text summarization is a summarization type where there is no addition of new content or modification of existing content, but rather only the most important phrases and sentences from the document are retained as the summary of document [1], [2]. It is akin to a highlighter used to highlight only the most important sections of a document to the viewer. Such an algorithm would require an accurate ranking of the sentences present in the document based on their relevance to the summary. Based on a decided threshold, the top N ranked sentences would then be predicted as the summary of the

document. Previous approaches in this domain have considered the use of statistical features such as word count, and term frequencies for ranking the sentences [3]. Traditionally, extractive text summarization algorithms have been demonstrated and conceptualized while taking into consideration the English language [4]. Recent years have seen a rise in the contributions from the research community for non-English languages, including many Indian languages [2]. Marathi is a language spoken in the state of Maharashtra and nearby regions in India and is derived from the Devanagari script [2]. The work on extractive text summarization in the Marathi language has been relatively paltry [5][6][7].

In this paper, two different approaches are proposed for performing extractive text summarization on Marathi text documents. The first approach focuses upon using latent semantic analysis (LSA) that deploys a semantic identification and correlation of sentences between a document [8]. This is achieved using the singular value decomposition (SVD) technique. The second approach makes use of fuzzy logic to solve this task. The fuzzy algorithm focuses on certain text characteristics and rules [9] using which sentence scores are assigned.

The contributions made through this paper are enlisted as follows:
1) Implementation of a latent semantic analysis based algorithm for extractive text summarization on Marathi documents
2) Implementation of fuzzy logic to derive rules useful for sentence ranking for extractive text summarization of Marathi documents.
3) Analysis and comparison of the aforementioned two approaches to better guide further work in this domain.
The rest of the content is structured as follows: previous work in this domain is discussed in Section 2, the two proposed approaches are described in detail in Section 3, the dataset description is detailed in Section 4, the results and analysis of the two approaches based on the results are carried out in Section 5 and the conclusion and future scope are mentioned in Section 6.

BACKGROUND AND RELATED WORK

Extractive text summarization has been the preliminary text summarization demonstrating the priority ranking capability of the data modeling algorithm. Previous work in this domain has focused on the use of statistical features in the early days, and recently has seen more focus on graphbased and deep learning-based methods.
The initial method for extractive text summarization made use of term frequency and inverse document frequency for feature selection [3]. Documents, however, could also include various themes that are addressed, and clustering these methods together was also used as an approach for priority ranking [10]. Kumar et al. [11] made use of a knowledge induced graph for performing singledocument summarization. In the last decade, machine learning has also been used to tackle this task across multiple domains as well as multiple languages [12], [13]. Query-based text summarization has also been tried where the ranking is based on the overlapping between the query phase and the document terms [14].

Recent years have also seen attempts to perform extractive text summarization using deep learning. Recently, there have been attempts to perform extractive text summarization in Marathi. Bhosale et al. [7] used a naive frequency count approach for this task. Rathod made use of the page-rank and text-rank algorithms, however, their evaluation of these approaches was very restricted [15]. Sarwadnya and Sonawane went a step further in terms of the use of preprocessing methods and reliance on the text-rank algorithm [5]. Chaudhari et al. [6] presented the use of deep learning to create the summarizer.

These approaches have either not been evaluated on a standard-sized dataset, or have made use of traditional approaches only. While, singular value decomposition and fuzzy logic have been explored in the English language [1], [16], there has been no significant contribution by the community for using these on Marathi documents. In this paper, an attempt is made to try to incorporate these approaches in a novel way to boost performance when working with Marathi documents.

PROPOSED METHODOLOGIES

Two different methods are proposed and analyzed in this paper- the first one is the use of singular value decomposition (SVD) strategy as a part of the latent semantic analysis (LSA) approach, while the second is the use of fuzzy logic and rules for deciding the ranking mechanism. Figure 1 shows the block diagram of the extractive text summarization architecture. Note that two different solutions are presented for the summarizer algorithm phase.

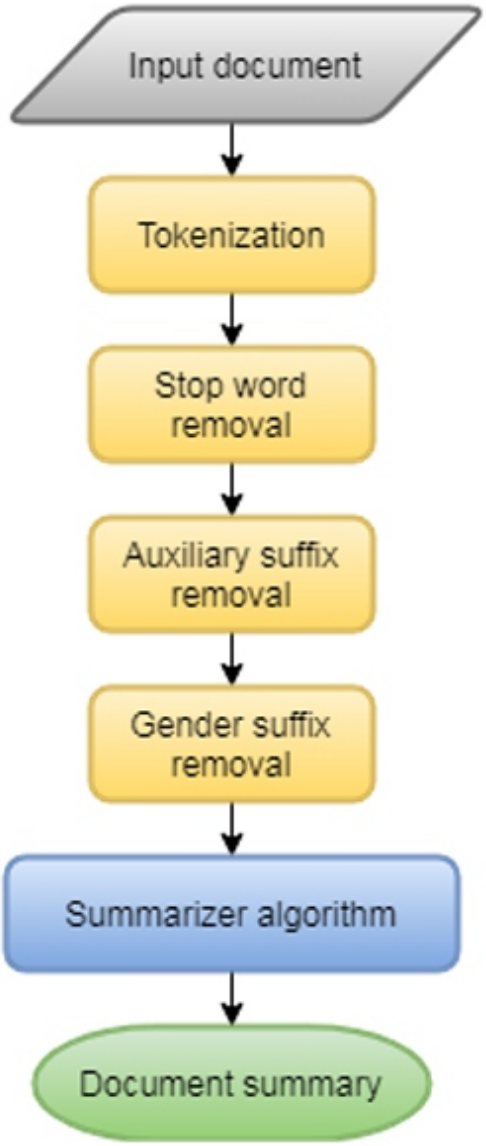

Figure. 1: Block diagram of the extractive text summarization architecture

Preprocessing

Preprocessing steps of the input documents in both the approaches remain the same. The document is first tokenized into separate tokens. This is followed by the removal of stop words. Stop words are the words that do not add to the meaning of the sentence and are used only to ensure the grammatical consistency of the sentence. These words do not add value in terms of realizing the ranking order of the sentences as they have a uniform probability of occurring in both important and unimportant sentences. Marathi language is characterized by the addition of suffixes to verbs to indicate the gender or the tense in which the sentence is being spoken. These suffixes also do not add any value to the semantic meaning of the sentence. They are removed to bring about faster processing and modeling and also reduce the number of distinct tokens modeled by the algorithm, thereby ensuring no ambiguous interpretations of similar meaning words. The preprocessed text is now more model-friendly and is passed as input to the summarizer algorithm.

Summarizer algorithms

Two different algorithms are presented in this paper for extractive text summarization. These are as follows:

1) Singular Value Decomposition (SVD): Computing the latent semantic structure of the document to obtain context similarity between the sentences and thereby mapping the vector space.

2) Fuzzy logic: Calculating values of some handcrafted statistical features and defining rules based on these features that are then passed as inputs to the fuzzy algorithm.

The inner working of both the algorithms is discussed in detail in the further subsections:

1) Singular Value Decomposition: Singular Value

Decomposition or SVD is a technique under latent semantic analysis that tries to correlate and find the relation between the sentences present in a document and the words present in that sentence. The approach works in two distinct phases: In the first phase, the input matrix D is created based on the term frequency of the words present in the document [17]. For m distinct words and n sentences in the document, D would be a mxn dimension matrix. As every word does not occur in each of the sentences, A tends to be a sparse matrix in nature. Further, every sentence row in this matrix is normalized to a range between 0 and 1 using the following equation:

$$sentence\_row = \frac{sentence\_row}{max(sentence\_matrix.value())} \quad (1)$$

Such a normalized input matrix can now be passed as an input to the SVD approach, which can be represented mathematically as follows:

$$D = U\Sigma V^T \quad (2)$$

Where, D: Normalized input representation matrix U: mxn matrix representing left singular vectors in the form of words x concept

Σ: nxn diagonal matrix indicating the singular eigenvalues, descending across the diagonal V : nxn matrix indicating the right singular vectors in the form of sentence x concept

Algorithm 1 indicates the procedure to derive the SVD values for subsequent ranking of the sentences in the document.

Algorithm 1 Algorithm for computing SVD

Input: Normalized input representation matrix D
Output: Values of U, V, and Σ
1: Prod_D = DD^T
2: x1 = Eigen_values(Prod_D)
3: Inv_D = D^TD
4: x2 = Eigen_values(Inv_D) 5: Val = $\sqrt{x}$——— 1 ∩ *x2*
6: Assign values to U, V, and Σ
7: **return** U, V, Σ

As a modification to the existing SVD approach, three other factors are also considered. Apart from the sentence similarity weight, the sentence length, sentence position, and the sentence value are also included to decide the final ranking for the summarization. Each of these factors is considered and evaluated as follows:

• Sentence length: If the sentence length is less than the minimum permissible length, or greater than the maximum permissible length, then set it to zero.

$$Sentence\ length = sin(\frac{180 * (Length -}{max\_length - min}$$

Otherwise, calculate as follows:

$$\frac{min\_length}{length(3)} \quad (3)$$

• Sentence value: The normalized input representation discussed earlier

• Sentence position: If the sentence is the first or the last one in the document, then consider it to be important and set value as 1. Otherwise, derive the value as follows:

$$Sentence\ pos = cos(\frac{Pos - (TRSH * len(sentences)) * 360}{(1 - 2 * TRSH) len(sentences)}) \quad (4)$$

Where, TRSH is a hyperparameter decided by the user. The value is set to 0.01 in the presented setup.

• Sentence weight similarity: Calculated using the number of overlapping words present between two sentences. The final ranking for the sentence is derived by considering the sum of the absolute values of each of these factors. Based on the summary factor given by the user, the ranked sentences are sorted in descending order and the filtered sentences are output as the summary of the document.

2) Fuzzy Logic: The proposed fuzzy logic is calculated using a feature matrix. The feature matrix is derived based upon certain statistical features present in the document. Each of these features is as follows:

• Position factor of the sentence: The position factor of the sentence is calculated by normalizing its order in the document with respect to the total number of sentences.

$$Pos\ factor = \frac{Total\_sentences - current\_pos}{Total\_sentences} \quad (5)$$

• Bigram token length: Bigram is the tokenization of words done, but by considering two words at a time. The number of such bigram tokens present in a sentence is considered.
• Trigram token length: Trigrams are similar to bigram, but they consider three words together at a time. Trigram token length refers to the number of such trigram tokens present in the sentence.
• TF-ISF vector: It considers the term frequency as well as sentence frequency and is calculated as follows:

$$tf\_isf = \frac{term\_freq}{sent\_freq * vocab\_pos} \quad (6)$$

• Cosine similarity: Calculate the cosine similarity of the sentence with respect to the centroid of the document. Mathematically, this can be represented as follows:

$$Cos\_similarity(S, Z) = \frac{S.Z}{||S||^2.||Z||^2} \quad (7)$$

Where, Z is the centroid of the document and S is the sentence in consideration.
• Thematic number: It takes into consideration the factor of the number of keywords present in a sentence with respect to the total keywords present in the document [9].

$$Thematic\_number(S) = \frac{keywords\ in\ S}{total\ keywords} \quad (8)$$

• Sentence length factor: It is calculated by taking the ratio of the length of the sentence to the length of the longest sentence present in the document [18].
• Numeric tokens: The number of numeric tokens present in the sentence in consideration with respect to its length.
• Pnoun score: The ratio of the number of proper nouns present in the sentence to the total words present in it. More important sentences generally tend to contain more information which would also be proportional to the number of proper nouns present in the sentence.

For each of these fuzzy variable factors, three values(poor, average, good) are used to auto-populate them. The fuzzy logic requires a triangular membership function generator that accepts an independent variable and a three element vector used to control the shape of the function [19]. Based on the previously mentioned nine factors, a consequent factor sent is determined that is termed as bad, average, and good for vector values of [0,0,50], [0,50,100], and [50,100,100] respectively. Using all of this information, five rules are set to compute the fuzzy logic prediction values. The rules are as follows:
1) sent['good'] = Position factor['good'] & Sentence length['good'] & Pnoun score['good'] & Numeric tokens['good']
2) sent['bad'] = Position factor['poor'] & Sentence length['poor'] & Numeric tokens['poor']
3) sent['bad'] = Pnoun score['poor'] & Thematic number['average']
4) sent['good'] = Cosine similarity['good']
5) sent['avg'] = Bigram token['good'] & Trigram token['good'] & Numeric tokens['average'] | TF-ISF['average']

For an instance of data, the values of the aforementioned nine factors are calculated per sentence and passed as input for the fuzzy logic to compute. If the output of the consequent factor is greater than 50, the sentence is included in the summary of the document.
Using these two methods, summarization of a standard size document dataset is carried out and the obtained results are discussed in the next section.

IV. DATASET DESCRIPTION

The performance of the two proposed approaches is evaluated on a custom created dataset consisting of Marathi news articles ranging on a diverse set of issues including politics, economics, and social affairs. The dataset consists of 100 documents coupled with their manual summaries used later for evaluation purposes. A sample instance from a document in the dataset is shown in Table I.

आर्थिक वर्षाच्या पहिल्याच दिवशी शुक्रवारी मुंबई शेअर बाजाराच्या निर्देशांकात (सेन्सेक्स) 72 अंशांची घसरण होऊन तो 25 हजार 269 अंशांवर बंद झाला. आशिया आणि युरोपीय शेअर बाजारातील घसरणीमुळे हा परिणाम झाला. दरम्यान, राष्ट्रीय शेअर बाजाराच्या निर्देशांकात (निफ्टी) 25 अंशांनी घसरून 7 हजार 713 अंशांवर बंद झाला. देशातील मोटारनिर्मिती क्षेत्रातील आघाडीची कंपनी मारुती सुझुकीच्या मार्च महिन्यातील विक्रीत 15.9 टक्के वाढ झाल्याने कंपनीच्या समभागात आज 0.11 टक्के वाढ नोंदविण्यात आली. कर्जाच्या विळख्यात सापडलेल्या जयप्रकाश असोसिएट्सने सिमेंट व्यवसायातील काही हिस्सा कुमारमंगलम बिर्ला यांच्या मालकीच्या अल्ट्रा टेक कंपनीला 15 हजार 900 कोटी रुपयांना विकण्याची घोषणा केली आहे. यामुळे जयप्रकाश continue

TABLE I: Sample document text

As both the methods are instance-based and do not involve any trainable parameters, the entire dataset consisting of all the 100 documents is used for evaluation purposes. The algorithm is predefined and hence segregation of data is not required with only one pipeline required for the entire task.

V. RESULT AND ANALYSIS

Extractive text summarization focuses on retention of the most important sections of the document. As a result, evaluation of such summarizers focuses on the amount of overlap between the human summary and the machine generated summary. To define this measure of overlap in a standard format, the ROUGE metric is used [20]. Given a human generated summary H and a machine generated summary M, the precision, recall, and the F1 score is defined as follows:

$$ROUGE1\ Precision = \frac{H \cap M}{M} \tag{9}$$

$$ROUGE1\ Recall = \frac{H \cap M}{H} \tag{10}$$

$$ROUGE1\ F1 = \frac{2 * R1.Precision * R1.Recall}{R1.Precision + R1.Recall} \tag{11}$$

Where ROUGE−1 refers to the overlap when considering unigrams i.e. one token at a time. In a similar manner, ROUGE-2 related metrics can be defined as follows:

$$ROUGE2\ Precision = \frac{bigrams\ in\ H \cap bigrams\ in\ M}{bigrams\ in\ M} \quad (12)$$

$$ROUGE2\ Recall = \frac{bigrams\ in\ H \cap bigrams\ in\ M}{bigrams\ in\ H} \quad (13)$$

The ROUGE2 F1 score is the harmonic mean of precision and recall. The ROUGE-L metric is defined similarly and refers to the longest matching subsequence amongst the two summaries [20]. Firstly, the performance of the approaches is evaluated on single-document summarization. Based on the mentioned performance metrics, the results are produced and tabulated in Table II.

TABLE II: Results obtained on both the approaches for single document summarization

Metric	SVD	Fuzzy Logic
ROUGE1:Precision	0.632	**0.641**
ROUGE1:Recall	0.623	**0.625**
ROUGE1:F1	0.612	**0.623**
ROUGE2:Precision	0.531	**0.561**
ROUGE2:Recall	0.519	**0.546**
ROUGE2:F1	0.512	**0.546**
ROUGEL:Precision	**0.665**	0.659
ROUGEL:Recall	0.614	**0.636**
ROUGEL:F1	0.626	**0.64**

It can be seen that the Fuzzy logic turns out to be a better approach as compared to the SVD method with better results on almost all of the performance metrics. Next, multidocument summarization is considered. In this case, the overlaps of tokens in the human summary and machinegenerated summary is considered across multiple documents. Evaluation is done for precision, recall, and the F1-score. The results obtained are shown in Table III.

TABLE III: Results obtained on both the approaches for multi-document summarization

Metric	SVD	Fuzzy Logic
Precision	**0.705**	0.625
Recall	**0.693**	0.655
F1	**0.682**	0.63

It can be seen that while SVD was lagging in case of single document summarization, it outperforms Fuzzy logic by a comfortable margin when it comes to multidocument summarization. Further, as a part of ablation studies, the F1 scores are compared with a recently used method of positionbased textrank [21].

While position-based rank seems to be the better performer for single document summarization, SVD turns out to be the better choice when working with multidocument summarization on Marathi documents. The results obtained for both the approaches on single-document

TABLE IV: Comparison of F1 scores with textrank algorithm

Metric	Position based Textrank	SVD	Fuzzy Logic
Multidocument F1	0.667	**0.682**	0.63
ROUGE1:F1	**0.646**	0.612	0.623
ROUGE2:F1	**0.592**	0.512	0.546
ROUGEL:F1	**0.659**	0.626	0.64

and multi-document summarization are visualized in Fig. 2 and 3. To analyze the findings, the advantages and limitations of both the proposed approaches in the case of Marathi language are noted in Table V.

TABLE V: Analysis of the two presented approaches

Approach	Fuzzy logic	SVD
Advantages	Can model nonlinear functions of arbitrary complexity [19] - Flexible and easy to implement - Based on natural language Accommodates faulty data	Performs better on multi-document summarization Acts as an initial step to advanced dimensionality reduction methods like PCA [8] Performs satisfactory approximation of data
Shortcomings	Does not consider the semantic analysis of the words Does not consider the correlation amongst sentences The number of rules keep on increasing with the number of input features	Might underperform on non-linear data - Fails to recognize the context and the meaning of polysemic words in the particular instance Semantic analysis is not performed

VI. CONCLUSION

In this paper, two novel approaches have been proposed for the task of extractive text summarization on Marathi documents. The first approach focused on singular value decomposition as a dimensionality reduction and feature selection technique. It took into consideration the sentence position and sentence length factors along with the calculation of eigenvectors. The second approach made use of fuzzy logic to derive rules used for priority ranking of sentences based on certain statistical features in the document. Both the proposed approaches have certain advantages and shortcomings. The evaluation of the approaches was done on a standard-sized dataset and a fuzzy logic-based approach was found to be better when working on single document classification. On the other

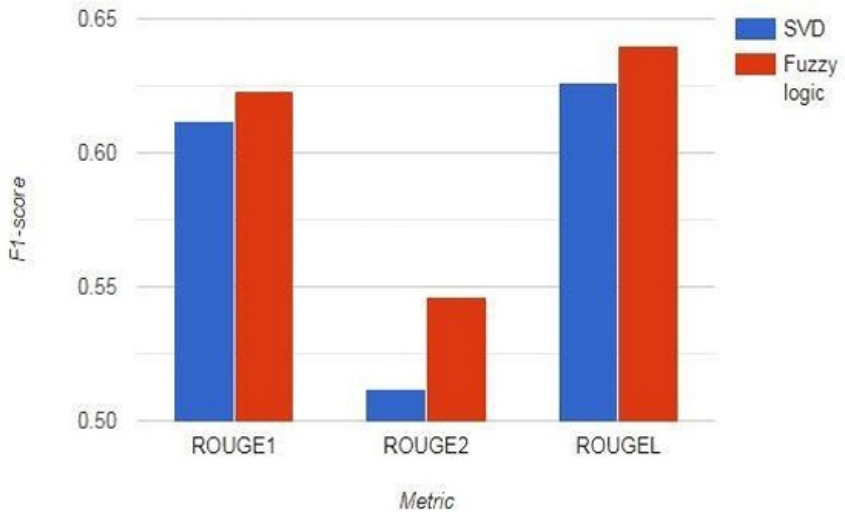

Figure. 2: Visual comparison of results obtained by both the approaches on single document summarization

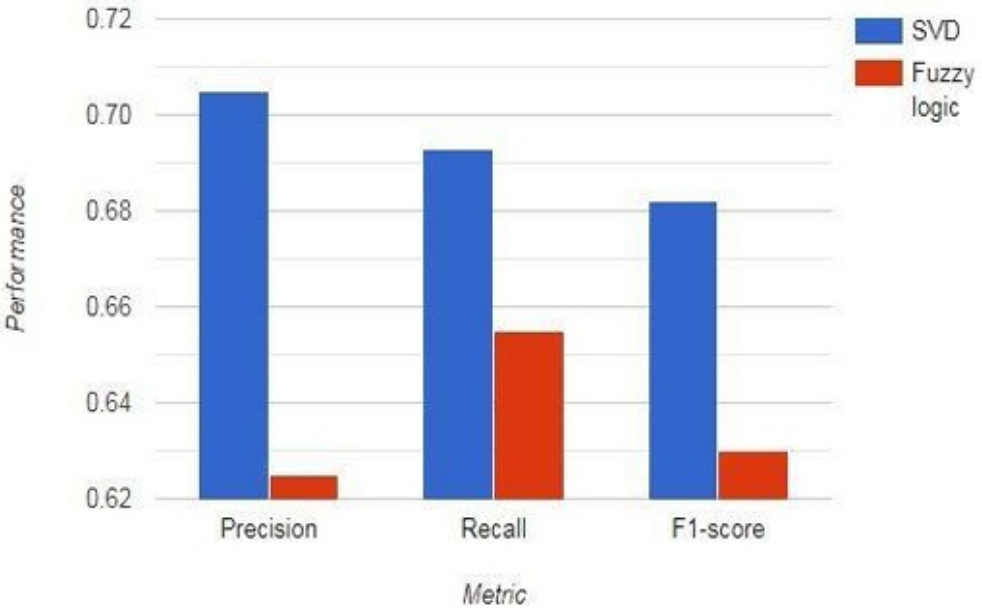

Figure. 3: Visual comparison of results obtained by both the approaches on single document summarization

hand, SVD was seen to be the better method for multi document summarization. There is a certain accuracy complexity tradeoff amongst the two approaches. This has been demonstrated by the evaluation of multiple performance metrics. As a part of ablation studies, the results were compared with another baseline method, and due analysis was carried out. Future scope in this domain includes consideration of semantic analysis, word embeddings, and extension of the task to include abstractive text summarization. Further, code-mixed text and low resource languages can be explored for this task. The proposed approaches have shown promising signs for text summarization task in Marathi language and could be extended further to other natural language understanding tasks.

REFERENCES

[1] V. Gupta and G. S. Lehal, "A survey of text summarization extractive techniques," Journal of emerging technologies in web intelligence, vol. 2, no. 3, pp. 258–268, 2010.

[2] V. V. Giri, M. Math, and U. Kulkarni, "A survey of automatic text summarization system for different regional language in india," Bonfring International Journal of Software Engineering and Soft Computing, vol. 6, no. Special Issue Special Issue on Advances in Computer Science and Engineering and Workshop on Big Data Analytics Editors: Dr. SB Kulkarni, Dr. UP Kulkarni, Dr. SM Joshi and JV Vadavi, pp. 52–57, 2016.

[3] J. Kaur and V. Gupta, "Effective approaches for extraction of keywords," International Journal of Computer Science Issues (IJCSI), vol. 7, no. 6, p. 144, 2010.

[4] K. Kaikhah, "Text summarization using neural networks," 2004.

[5] V. V. Sarwadnya and S. S. Sonawane, "Marathi extractive text summarizer using graph based model," in 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), pp. 1–6, IEEE, 2018.

[6] A. Chaudhari, A. Dole, and D. Kadam, "Marathi text summarization using neural networks," International Journal for Advance Research and Development, vol. 4, no. 11, pp. 13, 2019.

[7] M. S. Bhosale, D. Joshi, M. V. Bhise, and R. A. Deshmukh, "Automatic text summarization based on frequency count for marathi e-newspaper," 2018.

[8] R. M. Badry, A. S. Eldin, and D. S. Elzanfally, "Text summarization within the latent semantic analysis framework: comparative study," International Journal of Computer Applications, vol. 81, no. 11, pp. 40–45, 2013.

[9] J. Yadav and Y. K. Meena, "Use of fuzzy logic and wordnet for improving performance of extractive automatic text summarization," in 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 2071–2077, IEEE, 2016.

[10] A. R. Deshpande and L. Lobo, "Text summarization using clustering technique," International Journal of Engineering Trends and Technology, vol. 4, no. 8, pp. 3348–3351, 2013.

[11] N. Kumar, K. Srinathan, and V. Varma, "A knowledge induced graph-theoretical model for extract and abstract single document summarization," in International Conference on Intelligent Text Processing and Computational Linguistics, pp. 408– 423, Springer, 2013.

[12] K. Sarkar, M. Nasipuri, and S. Ghose, "Using machine learning for medical document summarization," International Journal of Database Theory and Application, vol. 4, no. 1, pp. 31–48, 2011.

[13] M. BAZRFKAN and M. RADMANESH, "Using machine learning methods to summarize persian texts," Indian J. Sci. Res, vol. 7, no. 1, pp. 1325–1333, 2014.

[14] I. Imam, N. Nounou, A. Hamouda, H. Allah, and A. Khalek, "Query based arabic text summarization 1," 2013.

[15] Y. V. Rathod, "Extractive text summarization of marathi news articles," 2018.

[16] P. D. Patil and N. Kulkarni, "Text summarization using fuzzy logic," International Journal of Innovative Research in Advanced Engineering (IJIRAE), vol. 1, no. 3, pp. 42–45, 2014.
[17] J. Steinberger and K. Ježek, "Text summarization and singular value decomposition," in International Conference on Advances in Information Systems, pp. 245–254, Springer, 2004.
[18] L. Suanmali, M. S. Binwahlan, and N. Salim, "Sentence features fusion for text summarization using fuzzy logic," in 2009 Ninth International Conference on Hybrid Intelligent Systems, vol. 1, pp. 142–146, IEEE, 2009.
[19] F. Kyoomarsi, H. Khosravi, E. Eslami, P. K. Dehkordy, and A. Tajoddin, "Optimizing text summarization based on fuzzy logic," in Seventh IEEE/ACIS International Conference on Computer and Information Science (icis 2008), pp. 347–352, IEEE, 2008.
[20] C.-Y. Lin and F. Och, "Looking for a few good metrics: Rouge and its evaluation," in Ntcir Workshop, 2004.
[21] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in Proceedings of the 2004 conference on empirical methods in natural language processing, pp. 404–411, 2004.

Fake News Detection Techniques for Diversified Datasets

Dr. M. Gayathri 1*, S. Tarini 2, S. Geetha 3
1, 2, 3 Department of CSE, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya, Kanchipuram, India

ABSTRACT

The introduction of the World Wide Web and the quick abandonment of the social media policy cleared the method for the rapid dispersal of information that has never been seen during human archive. Due to the way social media manifesto are currently operating, users are producing and participating in more information than ever before, some of which is false and has no relevance to reality. The numerous lives of individualities now hang in the balance as a result of social media. important has formerly been fulfilled in these three fields, including contact, advertising, news, and docket advancement. Automated bracket of a textbook composition as misinformation or intimation is a grueling task. Indeed, an adept in a distinctive sphere must traverse multiple features before granting a decree on the probity of a composition. In this work, we bring forward to use a machine literacy quintet perspective for the automated bracket of newspapers. [1] Our study traverses contrasting textual parcels that can be used to discriminate fake appease from real.

Social networking is one of the most critical subjects in the business world moment. For that reason, it is critical to pinpoint a vicious account. So, for that purpose we have developed machine learning algorithms to declare the real or fraud news. Machine learning algorithms will give the impose information about the data sets. These algorithms can decide to corroborate the real or fake news. [2] We have developed seven algorithms so that because of using these many algorithms finally we can compare the accuracy of all the algorithms. So, it can be tranquil to declare about the social media news. The data has been anatomized for these purposes, and learning algorithms have been used to identify fake news. By using these parcels, we instruct a coalescence of dissimilar machine learning programs using colorful septet styles and estimate their presentation on real world data files. Investigational appraisal confirms the supercilious presentation of our proposed chorus beginner perspective in correlation to solitary novice.

Keywords Artificial Intelligence, Authenticity, Classification, Fake News, social media, Websites.

INTRODUCTION

Numerous sociological studies that highlight the collision of fraud news and how people react to it have drawn the attention of many scholars in recent months. This phenomenon is known as fake news discovery (FND). One must first define fraud news before defining phony news as any content competent of leading readers to trust in information that is untrue. Spreading false information widely is bad for both the individual and community. Originally, this type of false news carried the threat of altering or upending the ecosystem's equilibrium of veracity. People are compelled to embrace false or biased ideas that they would otherwise reject because of the attributes of false news. The use of fraud news and propagandists is frequently used to convey political dispatches or impact. Fake news continues to affect how people react to and engage with real news. It is essential to create a system that can accordingly identify wrong news when it surfaces on social media in order to lessen its potentially dangerous effects. Still, there are several sensitive problems with uncovering fake news on various social

media platforms. [3] A variety of exploration objects configured in this regard includes the recognition of the source of origin or exchanging of the news or data on the social network, to understand the factual intention or meaning of the data uploaded and to determine the extent of legitimacy and validate it to make decision to consider it as genuine or fake. Automated false news detection is challenging because of news tricks. The lack of sufficient supporting claims or data prevents knowledge bases from successfully validating fake news when it is connected to time-critical programs.

False news also generates big, noisy, untreated data that is present on social media. In recent years, experimenters have attempted to recognize issues with fake news, specifically their accountability on social media, particularly Twitter, YouTube, Facebook, and TV. Because of these webbing connections, it is feasible to value important post columns while also utilizing the connections within the network. These characteristics, types, and discovery methods of fake news are all discussed in this research.

MATERIALS AND METHODS

Existing System

There takes place a vast body of study on the content of machine literacy styles for news discovery, utmost of it has been concentrating on classifying online critiques and openly available social media posts. The main problem of pinpointing fake news has received notice in the writings, extremely since late 2016 during the American Presidential election. Outlines numerous approaches that feel promising towards the end of fully classify the false papers. [4] They mention that easy content linked n- grams and shallow corridor part- of- speech trailing has demonstrated inadequate for the bracket work, frequently lacking to regard for dominant environment information, these styles have been shown precious only in cooperation with further complex ways of unifications.

Proposed Method

Proposed system because of the convolution of fraud news discovery in social media, it is apparent that a doable system must repress specific exposure to directly attack the issue. thus, the proposed system is a merger of semantic analysis. The proposed system is completely collected of Artificial Intelligence perspectives, [5] which is expository to directly relegate between the real and the untrue, rather of using algorithms that are unfit to empirical functions. The three- part system is a combination between Machine Learning algorithms that divide into natural language processing styles. Although each of these propositions can be merely pre-owned to classify and descry false news, in order to extend the delicacy and be germane to the social media sphere, they have been combined into a supervised machine learning algorithm [6] as a system for fake news discovery. It is important that we've some medium for detecting fake news, or at the veritably least, a mindfulness that not everything we read on social media may be true, so we always need to be allowing critically. This way we can help people make further informed opinions and they will not be wisecracked into allowing what others want to manipulate them into believing.

Architecture

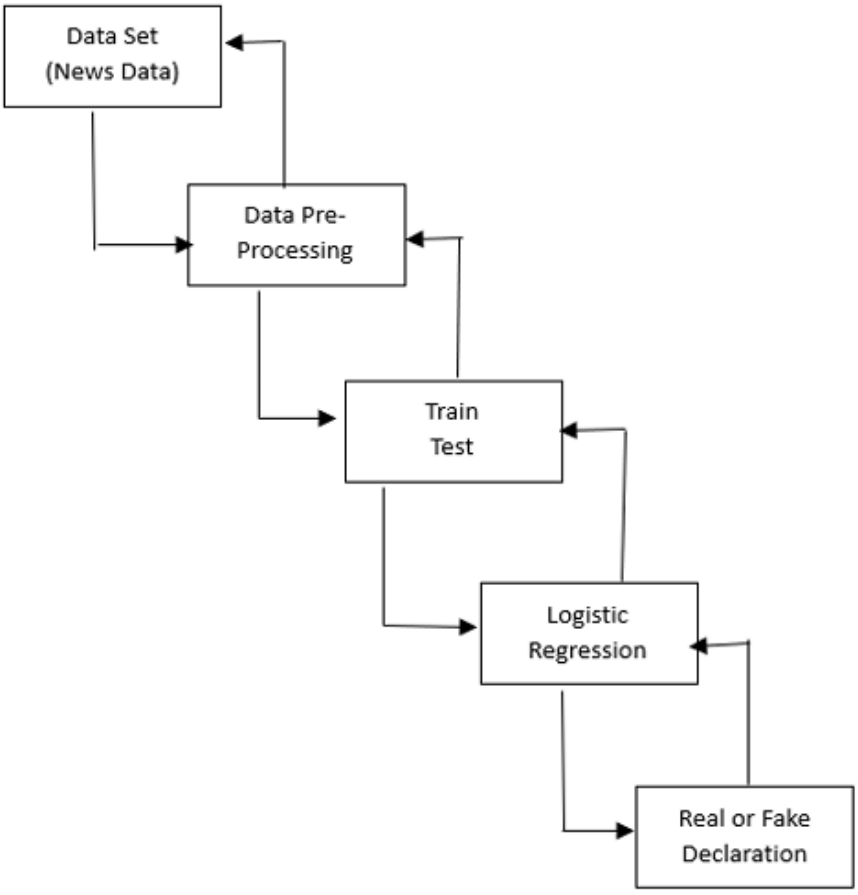

Figure 1. Architecture of Data Processing

According to the architecture shown in figure 1, the data will be gathered from the social media. And the data should be particularly regarding the social media news. The data should be especially news data. The gathered news data should be pre-processed for the further process. In this news data 70% of the data will be sent to training and the rest 30% of the data from the news data will be sent to testing. In every machine learning algorithm, the same process will be done. Then the seven algorithms are used to process the data. The data will be processed clearly in the machine learning algorithms. Then the algorithms will declare the news whether it is real or fake. [7]

Algorithms

TFIDF Vectorizer

TFIDF, short for term frequency – inverse document frequency, is a fine dimension which is conscious of how significant a expression is to a record in a multifariousness or aggregation. It is regularly employed as a weighting factor in quests of data recovery, textbook mining, and customer displaying. [8]

Logistic Regression Classifier

The probability of an objective variable is predicted using the supervised literacy bracket algorithm known as logistic regression. There are only two possible classes because the dependent term is dichotomous in character. Simply put, the dependent variable is a double with data encoded as either a 1 (for success/yes) or a 0 (for failure/no). A logistic retrogression model forecasts P (Y = 1) as a function of

X numerically. It is one of the most important straightforward ML algorithms that can be applied to issues with colored brackets like spam discovery, diabetes vaticination, cancer discovery, etc. Although logistic regression typically refers to double logistic regression with double target variables, it is also capable of predicting two additional levels of target variables. [9]

Decision Tree Classifier

Although Decision Tree algorithm is a supervised literacy approach which can be likely used for both bracket and regression problems, it is primarily favored for answering bracket problems. It is a tree-structured classifier in which the interior bumps stand in for the dataset's attributes, branches for the decision directives, and each splint knot for the outgrowth. The Leaf Node and the Decision Knot are the two peaks in a decision tree. Decision bumps are used to make the decisions and have multitudinous branches, considering that Leaf bumps are the subject of those views and do not have any additional branches. Using the characteristics of the provided information as a foundation, opinions or tests are conducted. For an issue, it is a graphical representation of all outcomes that could be achieved. [10]

Random Forest Classifier

Popular supervised reading algorithm Random Forest is part of the machine literacy movement. It can be applied to ML Bracket and Regression issues. The Random Forest classifier, as its name suggests, averages the results from various decision trees applied to vivid regions of the input dataset to diminish the delicate forecasting of the dataset. The arbitrary timber receives the vaccination from each decision tree and bases its prediction of the result on the maturity votes of prognostications rather than depending solely on one tree. due to the lack of vegetation trees in the wood, it is more delicate and the overfitting issue is avoided. [11]

SVM (Support Vector Machine) Classifier One of the most well-understood algorithms for supervised literacy, called Support Vector Machine (SVM), is used to solve Bracket and Regression issues. Nevertheless, it is mostly employed for Machine literacy bracket issues. The impetus of the SVM algorithm is to construct a chic line or decision boundary that can divide an n-dimensional space into groups so that new data points can be easily appended in the following process and placed in the actual order. A hyperplane is the title of this chic judgment boundary. SVM selects the extreme points that support in the construction of the hyperplane. Support vectors are what are mentioned to as these extreme instances, which is why the algorithm is named in this way. [12]

Naive Baye's

A batch of bracket algorithms invigorate on the Bayes' Theorem make up naive Bayes classifications. It is a collection of algorithms as a substitute of a singular algorithm, and they all share the same directing principle—namely, that each pair of hallmarks being divided is independent of the others. This algorithm, which is employed in a variety of machine literacy issues, operates on the Bayes theorem under the presumption that it is free from predictors. In other words, Naive Bayes works under the premise that each function in the sequence is independent of the others. [13]

Passive Aggressive Classifier

The Passive- Aggressive algorithms are the part of the machine learning procedures that are not well understood by learners and even intermediate Machine Learning tools. However, they can still be genuinely helpful and systematic for some tasks. An explanation of the algorithm's operation and appropriate applications is provided in this high-level summary. The principles behind how it functions

are not covered in detail. For widespread reading, passive-aggressive algorithms are usually used. One of many "online literacy algorithms" exists. As opposed to batch machine learning, which uses the complete training dataset all at once, online machine learning algorithms streamline the machine literacy model step-by-step.

This is useful in situations where there is a more quantum of data and it is mathematically infeasible to train the entire dataset because of the utter size of the data. We can normally say that an online- literacy algorithm will get a training demonstration, modernize the classifier. Passive- Aggressive algorithms are called since.

Passive: If the vaticination is true, maintain the model and do not make any substitutes. i.e., the data in the illustration is not abundant to beget any commutes in the model.

Aggressive: If the vaticination is false, make substitutes to the model. i.e., some alternatives to the model may correct it. [14].

RESULTS

Accuracy levels of the Algorithms

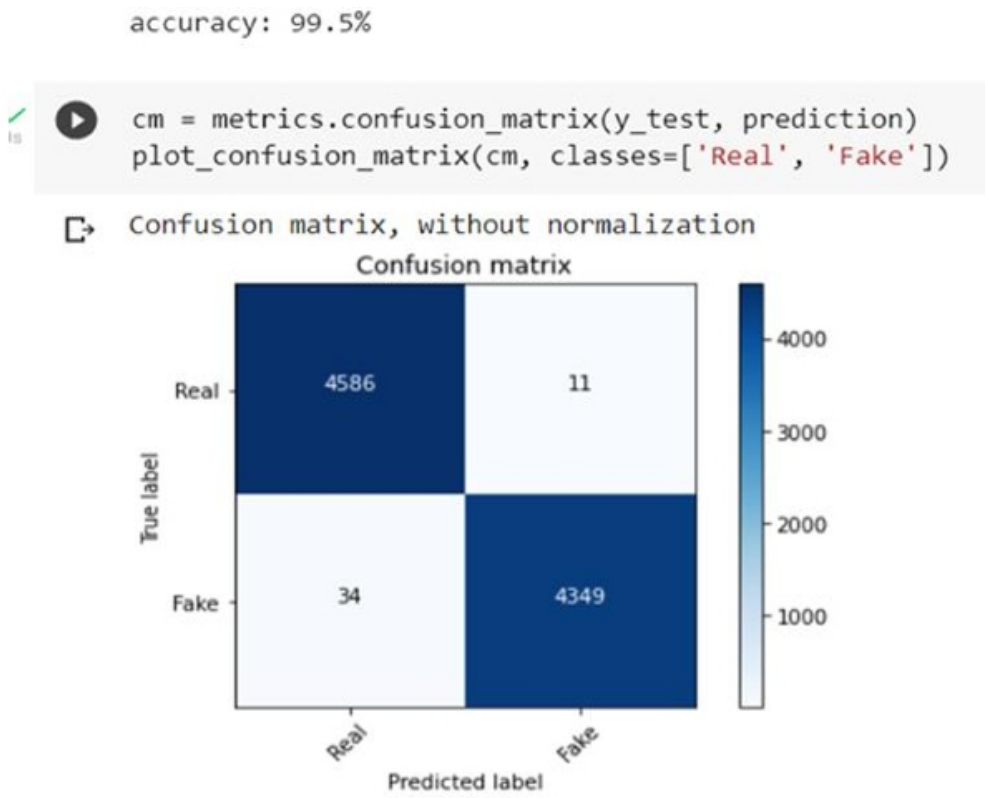

Figure 2. Accuracy of the algorithms

Output of all Algorithms

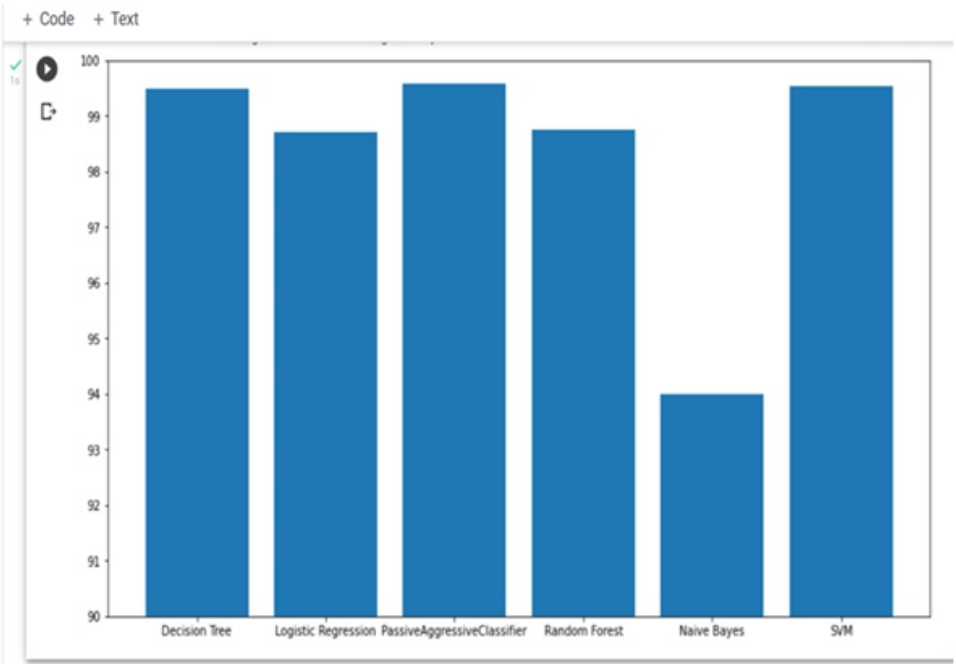

Figure 3. Graph of all algorithms

Comparison of all Algorithms

Table 1. Comparison of the algorithms

S.NO.	ALGORITHM	ACCURACY
1.	Decision Tree Classifier	99.6%
2.	Logistic Regression Classifier	98.76%
3.	Passive Aggressive Classifier	99.6%
4.	Random Forest Classifier	98.69%
5.	Naïve Bayes Classifier	94.18%
6.	Support Vector Machine	99.47%

DISCUSSION

Additional people now get the most of the information from social media than from the outdated fourth estate as a result of social media's improving content. Social media is also frequently given to straighten out fraud news, which has detrimental out-turn on both solitary consumers and society at large. By reviewing the available writings in two phases- depiction and detection, we often explore the drawback of fake news. [15] We presented the fundamental approaches and tenets of fake news in both customary and social media during the depiction section. In the detection section, we mustered existing false news detection techniques from a knowledge mining viewpoint, along with characteristic parentage and model building. We also tend to discuss the datasets, analysis metrics, and encouraging subsequent paths

in fake news detection and swell the sphere to subsequent appeals.

CONCLUSION

Taking the help of these various machine learning algorithms similarly Logistic Regression Algorithm, Tfidf Vectorizer, Decision Tree Algorithm, Random Forest Algorithm, SVM classifier, Naive bayes classifiers, Passive aggressive Classifiers etc. We have developed a model to forecast the news we took is a "True news" or "Fake news." Moreover, each classifier's results are successful. Some of them give the best results which have more accuracy, some of them have low accuracy. We are choosing the best of these models so that the results of the model will have more accuracy and give the more accurate results for the models.

Because of the overuse of social media, many individuals gather news from social media rather than olden methodologies. social media has conjointly been customary unfold affected news, that has sturdy bad impacts on individual users and wider society. We tend to traverse the affected news drawback by assessing present literature in two phases depiction and detection. Within the depiction part, we tend to introduced the required ideas and postulates of faux news in each earliest media and social media. Within the detection part, we have proclivity to evaluated existing pretend news detection approaches from a knowledge mining viewpoint, together with hallmark extraction and model construction. We have proneness to conjointly more addressed the datasets, survey metrics, and favorable subsequent directions in profess news detection analysis and spread the sphere to unconventional implementations.

As we can conclude that if we use the lower size data we get the accuracy results low as we use the larger size data set we get the results with more accuracy with these data we can have the Decision tree with higher accuracy as per the present dataset.it will change the accuracy according to the dataset, the second highest accuracy shown in the plot is SVM classifier but it takes more time than the other algorithms so we consider the third highest accuracy which gives the good results for our model as show in the plot diagram we consider Passive aggressive classifier as the best algorithm for our model.

REFERENCES

[1] The Journal of Supercomputing, vol. 76, no.7, pp.4802–4837, 2020. K.S. Adewole, T. Han, W. Wu, H. Song, and A.K. Sangaiah. Twitter spam account detection based on clustering and classification methods. https://link.springer.com/article/10.1007/s11227-018-2641-x

[2] Soft Computing, vol. 24, no. 5, pp. 3475–3498, 2020. M.Z. Asghar, A. Ullah, S. Ahmad, and A. Khan. Opinion spam detection framework using hybrid classification scheme. https://link.springer.com/article/10.1007/s00500-019-04107-y

[3] International Journal of Multimedia Information Retrieval, vol. 7, no. 1, pp.71–86, 2020. C.Boididou, S. Papadopoulos, M. Zampoglou, L. Apostolidis, O. Papadopoulou, and Y. Kompatsiaris. Detection and visualization of misleading content on Twitter. https://link.springer.com/article/10.1007/s13735-017-0143-x

[4] H. Dathar Abas, and A. Mohsin Abdulazeez. "A Modified Convolutional Neural Networks Model for Medical Image Segmentation." learning 20 (2020).

[5] Brooks, Gabriel. "Introduction to Python Pandas for Beginners". Almabetter.com. Retrieved 24 October 2020.

[6] K. Shu, S. Wang, and H. Liu. "Exploiting tri-relationship for fake news detection." arXiv preprint arXiv:1712.07709 8 (2017).

[7] Uma Sharma, Sidarth Saran, Shankar M. Patil. "Fake News Detection using machine learning algorithms." Machine learning IJCRT (2020).

[8] "NumFOCUS Sponsored Projects". NumFOCUS. Retrieved 2021,10-25.

[9] C. Zhou, et al. "Boost classifier for DDoS attack detection and analysis in SDN-based cloud." 2018 IEEE international conference on big data and smart computing (big comp). IEEE, 2018.
[10] International Journal of Machine Learning and Cybernetics, vol. 10, no. 8, pp. 2143–2162, 2021. K. Dhingra and S.K. Yadav. Spam analysis of big reviews dataset using Fuzzy Ranking Evaluation Algorithm and Hadoop. https://link.springer.com/article/10.1007/s13042-017-0768-3
[11] D. Longjun, et al. "Discrimination of mine seismic events and blasts using the fisher classifier, naive Bayesian classify and logistic regression." Rock Mechanics and Rock Engineering 49.1 (2016), 183-211.
[12] IJERT-Fake News Detection using Machine Learning Algorithms. Uma Sharma, Sidharth Saran, Shankar M. Patil. https://www.academia.edu/download/66254531/fake_news_detection_using_machine_IJERTCONV9IS03104.pdf.
[13] Abdulqader, Dildar Masood, Adnan Mohsin Abdulazeez, and Diyar Qader Zeebaree. "Machine Learning Supervised Algorithms of Gene Selection: A Review." Machine Learning 62.03 (2020).
[14] M. Granik, Mykhailo, and V. Mesyura. "Fake news detection using naive Bayes classifier." 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON). IEEE, 2017.
[15] S. Helm Stetter, and H. Paul Heim. "Weakly supervised learning for fake news detection on Twitter." 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 2018.

Feature Selection using Random Forest Classifier for Foot Strike Event Detection in Toe Walkers

Meghna Desai1*, Dr. Viral Kapadia2
12Department of Computer Science and Engineering, Faculty of Technology and Engineering, The Maharaja Sayajirao University of Baroda, Vadodara, Gujarat, India

ABSTRACT

Automated Gait event identification of Foot Strike (FS) and Foot Off (FO) in pathological gait data, can be time saving in comparison to conventional manual annotations done currently. Identification of FS and FO allows breaking walking trials into gait cycles and hence aids in comparison of gait parameters like joint angles, forces and moments across gait cycles. Automated Gait Event Detection is also useful in development of wearable sensor devices and robotic systems that assist gait. Researchers have proposed several automatic gait event detection algorithms based on kinematic parameters and systematic study of the literature suggests specific parameters to have higher contribution in identification of FS event in all common pathological gait patterns. We used Random Forest Classifier Feature selection technique to identify high contributing features in FS event in toe walking pediatric pathological gait dataset and the results suggest high similarity in selected features by the machine learning technique with those suggested by popular event detection algorithms based on kinematic parameters for pathological gait. Hence we conclude that RFC feature selection is suitable for feature selection in toe walkers gait dataset for event detection purpose.

Keywords : Feature selection, foot off, foot strike, pathological gait.

INTRODUCTION

Gait refers to a person's manner of walking. Normal gait is a repeated cycle of rhythmical, alternating movements of the body which results in its forward movement [1]. Normal gait consists of two phases. These phases are further divided into a total of 8 subphases. The first subphase of a normal gait cycle is called stance phase, it occupies 60% of the complete gait cycle during which some part of the concerned foot is in contact with the ground. The further division of stance phase is done into Initial Contact (foot / heel strike), loading response (foot flat), Midstance, terminal stance, Pre-swing (toe off/ foot off). The second subphase of a normal gait cycle is called swing phase, it occupies 40% of the total gait during which the concerned foot is not in contact with the ground and the body weight is borne by the other leg and foot. The further division of swing phase is done into Initial swing, mid swing and late swing.

Pathological gait is an altered gait pattern which can occur due to deformities in limbs, weakness, injuries, ageing or medical conditions like cerebral palsy, parkinsons disease, stroke, multiple sclerosis or other impairments. Gait abnormality can have tremendous impact on the patients especially on the quality of life, can cause severe injuries [2].

Gait analysis is an assessment of the way a person walks or runs from one place to another. 3d Gait analysis is done in gait laboratories for people with impaired gait, especially CP children. The results of the 3D Gait analysis are used to diagnose gait issues, track the progression of disease, measure the improvement in gait due to intervention/ rehabilitation/ therapy given to patient. Gait Event Detection is

essential for gait analysis. During gait analysis the gait variables at joint angles, forces and moments observed at specific events during a gait cycle are compared, so gait cycle determination becomes a primary task. Gait cycles can be determined from walking trials by the detection of Initial Contact (IC/FS/HS) and toe off/Foot Off (TO/FO) events. However, Gait Event detection is a highly time-consuming process in 3d Gait Analysis [3] [4]. Force plate measurements calculated from ground reaction forces are considered the gold standard in the task of gait event detection [3] [4] [5]. Force plates are not always installed in gait laboratories and in case of pathological gait like CP it is not always applicable as force plate strikes may not be clear and that results in false force thresholds many times. The cost of installation and maintenance of force plates in the laboratory restricts the number of cycles available for measurement and in pathological or pediatric gait clean force plate hits may not be possible due to simultaneous multiple steps on same force plate or if gait is assisted by devices like croucher or walker [6] [7]. In the case of pathological gait, manual gait event detection of IC and FO is required, which is time consuming and can result in human error due to visual inspection of gait events.

Accurate and efficient automated gait event detection can make gait analysis process comparatively fast and error free, aid in calculating spatio temporal parameters and is also required for development of wearable sensor devices and robotic systems that assist gait. The different types of quantitative data collected/computed during gait analysis includes kinematic, kinetic, oxygen consumption electromyography. Kinematic parameters of walking gait include displacement of the body, orientation of the body, joint angles and spatio-temporal data. Most of the automated gait event detection (AGED) algorithms are based on kinematic parameters [3] and this paper will also focus on automated gait event detection methods based on kinematic parameters. For performance evaluation most of the gait event detection algorithms use either force plate data (if available) or manual identification of gait events performed through visualization of markers trajectory as ground truth data. The literature review suggests some highly important kinematic features in detecting IC in pathological gait patients. In this paper we apply machine learning based feature selection technique and check its applicability to feature selection for kinematic gait data obtained from 3d instrumented gait analysis by comparing the same with information derived from literature review.

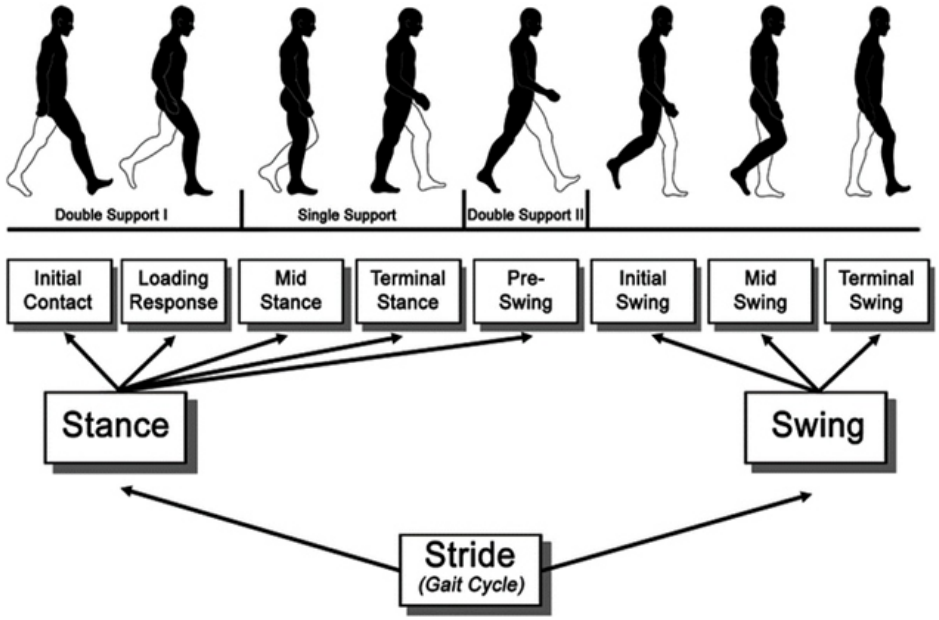

Figure 1. Gait Cycle phases and sub phases according to [8]

LITERATURE SURVEY

Researchers have used different kinematic parameters and proposed algorithms for IC detection, most of which show good accuracy for normal gait [6]. Comparing the performance of different algorithms, which use different kinematic parameters for gait event detection, on same pathological dataset can provide a basis for comparison, determination and recommendation of the most suitable technique. Researchers [3] [5] [9] have compared these algorithms on pathological datasets of subtle sizes and based on those results recommended the approach that can be used for AGED in pathological gait. [3] identified four gait patterns and classified each child participant in one of the patterns, then compared the results obtained by implementing nine published kinematic AGED algorithms [10] [11] [12] [13] [14] [15] [16] [17] [18] on a pediatric gait database (primarily CP pathologies) with more than 750 total manually annotated events. For FS they recommended the kinematic features sagittal resultant velocity [12], horizontal position [11] [18] or vertical/horizontal acceleration [13] [14] depending on whether the participant's terminal swing was observed to be more horizontal or vertical. For TO/FO, their recommendation was horizontal position [11] [18] and Sagittal Velocity [12] for all classified gait patterns. They also recommended algorithm determined by [12] in case when only one algorithm was preferred in common for IC and FO event detection across all identified gait patterns.

Another research classified the participants into 3 gait patterns namely Toe walkers, Flat IC and Heel IC, and compared the results obtained by implementing five kinematic AGED algorithms (one modified) [11] [12] [16] [18] on pediatric gait dataset of 90 children which was already rated with visual and force-plate mechanisms[9]. The recommendations given for IC included Sagittal Velocity of the heel for Heel IC pattern and Sagittal Velocity of the toe marker configurations for Toe Walkers and Flat IC groups [9]. Sagittal velocity of the hallux marker configuration for FO/TO was also recommended [3].

One more study classified seven CP participants in 2 gait patterns and collected kinematic and kinetic data for a total 202 steps with 202 FS and 194 FO events detected using force plate [5]. The FS and FO events were detected by implementing five AGED algorithms [11] [12] [13][14] [18] on this dataset and the results were compared with those obtained by the detection of these same events using the force plate. They concluded that AGED algorithm for IC and FO algorithm determined by [12] was recommended in children with Spastic Cerebral Palsy (SCP) when force plates were not available.

Recently researchers have also applied machine learning and deep learning techniques for AGED in pathological patients. In one study the researchers trained a multilayer feed forward neural network using the kinematic data obtained from cohort of 50 pathological subjects from which 29 walked barefoot and 21 shod/braced [19]. They used kinematic parameters sagittal plane position, velocity and acceleration of the heel and toe markers, foot-floor angle, angular velocity and angular acceleration to describe each frame of motion capture data. PCA was applied for dimensionality reduction. The trained multilayer feed forward neural network's event detection method was validated using kinematic data of 40 pathological patients. The comparison of results obtained from the neural network method with that of ground truth results obtained from force plate was in agreement within 1 to 2 frames in most of the cases, which assured the applicability of neural networks trained using kinematic gait data for AGED task [19].

[6] used three-dimensional coordinate and velocity based kinematic parameters obtained from 3d gait analysis in a gait laboratory to train and validate an LSTM model for AGED of FS and FO. They used a pediatric pathological gait dataset consisting of 18153 walking trials with 9092 annotated FS events was used to train and validate the constructed LSTM model(s).The best performing model identified FS with

an average error of 10 milliseconds and FO events with an average error of 13 milliseconds. The applicability of deep neural networks for AGED using kinematic gait data was determined [6].
[20] used 3d position and velocity of markers on the toe and lateral malleolus to train and validate a bilateral LSTM for AGED of FS and FO. A pediatric pathological gait database of 226 children with 1156 trials having manually annotated gait events was used to train and validate the Deep Event recurrent network. They also compared the results of their deep learning model with the results from AGED obtained by implementing the same dataset on [18] [12] [14] [6] [20] and based on results obtained recommended their proposed deep event model for AGED of FS and FO.

EXPERIMENTS

When the heel is unable to contact the floor at the explicit beginning of stance phase or the absence of first heel rocker is defined as toe walking [21] [22] [23] [24][25]. Toe walking is observed to be a common disorder in hemiplegic children and diplegic children with Cerebral Palsy [21] [22] [23] [24][25]. Gait Analysis results in collection of high number of kinematic parameters. Systematic review of literature reveals specific kinematic parameters to have high contribution in identifying the FS event in gait cycle for all common gait pathological patterns.These features/parameters are listed in Table1. Gait event identification is basically a problem and machine learning, deep learning may be suitably applied for the same [6] [19] [20]. We attempt to carry out Feature selection using machine learning to the gait dataset because large number of kinematic parameters are collected from gait analysis. The purpose of this paper is to check the suitability of a well known feature selection technique Random Forest Classifier, to the paediatric pathological gait dataset collected from 19 toe walking patients.The resultant important features in order of ranks assigned by Random Forest Classifier are compared to features listed in Table 1.

Table 1

Significant Gait Kinematic Parameter in FS detection in Pathological Gait derived from Literature Survey
Sagittal Heel Velocity
Sagittal Velocity Toe 5
Toe and Heel Marker Longitudinal (Z Component) Position and Velocity
Linear Velocity of Heel Marker (X Component)

DataSet

A retrospective study was conducted on the dataset consisting of 23 kinematic features (listed in Table 2) from 115 walking trials of 19 patients from a paediatric toe walking gait analysis dataset. The dataset was determined from the 3d gait analysis of the patients carried out at gait laboratory Jupiter Hospital, Thane, India.

Experiment Details

Feature Selection was carried out using sci-kit learn library. The csv file containing the gait data consisted of 93422 frames marked with 590 FS events manually annotated by the laboratory engineer. Table 2 lists the ranks given to the features/ gait parameters by the feature selection algorithm. Figure 2

shows the plot of obtained feature importances using mean decrease in impurity.

Table 2. Ranks of Features by Random Forest Classifier

Column Number in dataset	Description of Parameter	Rank Given by Random Forest Feature Selection
Feature 0	Knee Angle X Component	10
Feature 1	Knee Angle Y Component	21
Feature 2	Knee Angle Z Component	7
Feature 3	Linear Heel Velocity X Component	9
Feature 4	Linear Heel Velocity Y Component	8
Feature 5	Linear Heel Velocity Z Component	6
Feature 6	Linear Toe 5 Velocity X Component	3
Feature 7	Linear Toe 5 Velocity Y Component	19
Feature 8	Linear Toe 5 Velocity Z Component	14
Feature 9	Linear Toe 2 Velocity X Component	4
Feature 10	Linear Toe 2 Velocity Y Component	15
Feature 11	Linear Toe 2 Velocity Z Component	20
Feature 12	Linear Toe 1 Velocity X Component	16
Feature 13	Linear Toe 1 Velocity Y Component	17

Feature 14	Linear Toe 1 Velocity Z Component	12
Feature 15	Sagittal Velocity Heel	1
Feature 16	Sagittal Velocity Toe 5	2
Feature 17	Sagittal Velocity Toe 2	11
Feature 18	Sagittal Velocity Toe 1	13
Feature 19	Vertical Acceleration Heel (Z Component)	5
Feature 20	Horizontal Acceleration Heel (X Component)	22
Feature 21	Jerk Heel (Z Component)	23
Feature 22	Jerk Heel (X Component)	18

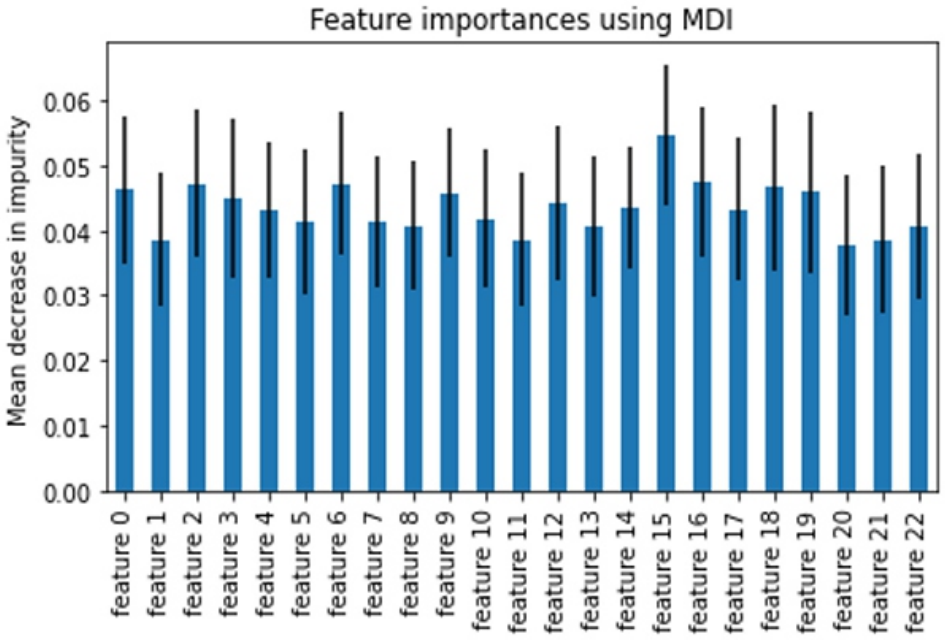

Figure 2. Feature importances using mean decrease in impurity

COMPARISION OF RESULTS

The most important features resulting from the input features given to the algorithm for feature selection are Sagittal Heel Velocity and Sagittal Toe 5 Velocity as per Table 2. The other important features in order are X Component of Linear Velocity of Toe 5 and Toe 2. Comparing these obtained results to the list of features/ gait parameters identified as important contributors in determining the FS from the research reviewed in literature, it is observed that Sagittal Heel Velocity , Sagittal Toe 5 Velocity are the common

parameters determined by both methods.

CONCLUSION

The experiments performed on the gait dataset of toe walkers and the analytical results achieved show a high similarity in the prominent features derived. The literature review has suggested Sagittal Heel Velocity and Sagittal Toe Velocity as important features and the same features have been rank highest by the feature selection technique using Random Forest Classifier for determining FS event in toe-walkers. Hence we conclude that random forest classifier feature selection technique suits the data of toe walkers. And the selected features can further be used to classify FS gait event in toe walkers using suitable algorithms.

ACKNOWLEDGEMENT

The authors are thankful to the Gait and Motion Analysis Laboratory, Jupiter Hospital, Thane for providing the data used in this research.

REFERENCES

[1] Esquenazi, & M. Talaty,. Gait analysis, Technology and clinical applications. Physical Medicine and Rehabilitation. pp.99-116, 2011.

[2] A.H.M Ata Ullah & O. Jesus, Gait Disturbances, Stat Pearls, NCBI Book Shelf, Jan 2022.

[3] D. Bruening & S.T. Ridge, Automated event detection algorithms in pathological gait, Gait and Posture, Elsevier, vol 39, Issue 1, pp. 472-477, January 2014.

[4] Gómez-Pérez, J. Martori, A. Josep M. Casanovas, J. Samsó, Josep M. Font-Llagunes, Gait event detection using kinematic data in children with bilateral spastic cerebral palsy, Clinical Biomechanics,Volume90,2021,105492.

[5] R.V. Goncalves, S. T. Fonseca, P. A. Araujo, V. L. Araujo, T.M. Barboza, G. A. Martins, M. C. Mancini, Identification of gait events in children with cerebral palsy: comparison between force plate and algorithms, Brazilian Journal of Physical Therapy, 2019. https://doi.org/10.1016/j.bjpt.2019.05.007.

[6] L. Kidzinski, S. Delp, M. Schwartz, Automatic real time gait event detection in children using deep neural networks, PLoS ONE 14(1): e0211466. https://doi.org/10.1371/journal.pone.0211466.

[7] Y.K. Kim, R M S Visscher, E. Viehweger, N. B Singh, W.R Taylor, F. Vogl, A deep learning approach for automatically detecting gait events based on foot marker kinematics in children with cerebral palsy-Which markers work best for which gait patterns?, PLoS ONE, October 2022.

[8] J. Perry and J. Burnfield, Gait Analysis: Normal and Pathological Function. SLACK Incorporated, 2010.

[9] R.M.S Visscher, S. Sansgiri, M. Freslier, J. Harlaar, R. Brunner, W. R. Taylor, N. B. Singh, Towards Validation and Standardization of automatic gait event identification algorithms for use in paediatric pathological populations, Gait and Posture, Elsevier, Vol 86, pp. 64-69, March 2021, https://doi.org/10.1016/j.gaitpost.2021.02.031.

[10] A.R. de Asha, M.A. Robinson, GJ Barton, A marker based kinematic method of identifying initial contact during gait suitable for use in real-time visual feedback applications, Gait and Posture, 2012:36(3):650-2.

[11] E. Desailly, D. Yepremian, P. Sardain, P. Lacouture, Foot Contact event detection using kinematic data in cerebral palsy children and normal adults gait, Gait Posture, vol. 29, pp. 76-80 ,June 2008.

[12] S. Ghoussayni, C. Stevens, S. Durham, D. Ewins, Assessment and Validation of a simple automated method for detection of gait events and intervals, Gait Posture, vol 20, pp. 266-272, 2003.

[13] A. Hreljac, RN. Marshall, Algorithms to determine event timing during normal walking using kinematic data, Journal of Biomechanics, 33(6):783-6, 2000.
[14] B-J Hsue, F Miller, F-C Su, J Henley, C Church, Gait timing event determination using kinematic data for toe walking children with cerebral palsy, Journal of Biomechanics, 2000.
[15] J M Jasiewicz, J HJ Allum, J M Middleton, A. Barriskill, P. Condie, B. Purcell, R.C.T Li, Gait event detection using linear accelerometers or angular velocity transducers in able-bodied and spinal-cord injured indviduals. Gait Posture, 2006.
[16] C.M. O'Connor, S.K. Thorpe, M.J.O Malley, C L Vaughan, Automatic detection of gait events using kinematic data, GaitPosture, 2007.
[17] Salazar-Torres J-D-J, Validity of an automated gait event detection algorithm in children with cerebral palsy and non-impaired children. GaitPosture, 2006.
[18] Jr. JA Zeni, JG Richards, JS Higginson, Two simple methods for determining gait events during treadmill and overground walking using kinematic data, GaitPosture, 2008.
[19] A. Miller, Gait event detection using multilayer neural network, Gait & Posture, 2009.
[20] M. Lempereur, F. Rousseau, O.R. Neris, C. Pons, L. Houx, G. Quellec, S. Brochard , A new deep learning-based method for the detection of gait events in children with gait disorders: Proof-of-concept and concurrent validity, Journal of Biomechanics, 2019.
[21] C. Beyaert, J. Pierret, R. Vasa, J. Paysant, and S. Caudron, Toe walking in children with cerebral palsy: a possible functional role for the plantar flexors, Journal of Neurophysiology, 2020.
[22] S. Armand, E. Watelain, M. Mercier, G. Lensel, FX. Lepoutre. Identification and classification of toe-walkers based on ankle kinematics, using a data-mining method. Gait Posture 23: 240–248, 2006. doi:10.1016/j.gaitpost.2005.02.007.
[23] M. Galli, E. Fazzi, F.Motta, M. Crivellini. Kinematic and dynamic analysis of theankle joint in children with cerebral palsy. FunctNeurol 14: 135–140, 1999.
[24] Rodda J, Graham HK. Classification of gait patterns in spastic hemiplegia and spasticdiplegia: a basis for a management algorithm. European Journal of Neurology 8, Suppl 5:98–108, 2001. doi:10.1046/j.1468-1331.2001.00042.x.
[25] TF. Winters Jr, JR Gage, R. Hicks, Gait patterns in spastic hemiplegia in children and young adults. Journal of Bone and Joint Surgery, Am 69: 437–441, 1987. doi:10.2106/00004623-198769030-00016.

A Novel Medical Chatbot with Alzheimer's Disease Detection Using Deep Neural Network

P.Maragathavalli 1*, Aishwarya Devi.V 2, Bhuvanesh.D 3, Manikandan.S 4
1 Assistant Professor, Information Technology, Puducherry Technological University, Puducherry, India
2, 3, 4 B.Tech Student, Information Technology, Puducherry Technological University, Puducherry, India

ABSTRACT

The healthcare sector is one of the largest focus areas in the world today. Individuals are becoming increasingly susceptible to lifestyle diseases. Hospitals and clinic are the most widely used place by the patients to consult doctor and get treated. People consider it as the most reliable means to check their health status. But in this way of approach for treatment the patients need to wait for a long time to consult the doctor which makes them more sick . In order to avoid such situation we came up with the idea of medical diagnosis chatbot in which user can interact with the Artificial Intelligence chatbot , to analyze the disease based upon the symptoms and with the MRI scan report.

Keywords : DNN, imagenet, inceptionV3, machine learning, mobilenet, MRI scan images.

INTRODUCTION

The main goal of the system is to create a medical chatbot in which the user can interact with the AI and diagnose the disease based upon the symptoms .The chatbot will be available at any time and the user can utilize it at any point of time when there is a requirement for it .Sometimes the chatbot may diagnose the disease wrongly because the symptom alone is not sufficient for it to analyse the disease In order to avoid such situation we came up with the idea of medical diagnosis chatbot in which the user can interact with the AI chatbot , to analyse the symptoms, risk factors and prevention of specific disease and can also detect the severity of Alzheimer's disease by uploading the MRI scan report. Here, we use the machine learning approach to built the chatbot.

MOTIVATION

The primary goal is to develop a prediction system which will allow the users to check whether they have Alzheimer Disease, the user need not visit the doctor unless he/she has Dementia or Alzheimer Disease, for further treatment.The secondary aim is to develop a web application that allows users to diagnose their disease based on the symptoms they have.This system will be available at any time and the user can utilize it at any point of time when there is a requirement for it.The accuracy of the prediction will be high and the time limit for the prediction will be low compared to the existing system.

LITEARATURE SURVEY

Uddin, M.Z., Dysthe, K.K., Følstad, has done research To automatically detect depression symptoms in text for decision support in health care is important. In this work, a multimodal human depression prediction approach has been investigated based on one-hot approach on robust features based on describing depression symptoms and deep learning method, RNN. Using the proposed approach,91% and 92% mean prediction performance has been achieved on datasets.It is only based upon the

prediction of depressive symptoms.It is based on only specific disease like depression and only has the text feature,designed in French language.
Arriba-Pérez, F., García-Méndez, S., González-Castaño, F.J. et al. has done research on Automatic detection of cognitive impairment in elderly people using an entertainment chatbot with Natural Language Processing capabilities.In this work, to reduce caregivers' effort and the whitecoat effect, we have proposed a novel conversational system for entertainment and therapeutic monitoring of elderly people.It relies on nlp techniques for chatbot behaviour generation and user-transparent automatic assessment, by combining ,distracting (user-centred) with attention-demanding questions (embedded cognitive tests).It achieved a accuracy of 90%.It is designed only for text feature for elder people in norwegian language. Junxiu Liu, Mingxing Li, Yuling Luo, Su Yang, Wei Li, Yifei Bi, has done research on Alzheimer's disease detection using depthwise separable convolutional neural networks, A novel DSC network-based method for detection of AD is pro- posed in this paper. The conventional CNN method is first used to detect AD, and the classification accuracy rate reached 78.02% in a three-way classification scenario (AD, MCI and normal). Then, an AD detection method combining
DSC and CNN is proposed. Compared with the CNN, the model parameters of the proposed method are reduced by 87.94% and the computing cost is reduced by 84.25%, where the classification accuracy rate remains moder- ately the same.It has only image feature and low rate of accuracy.

Arjaria, S.K., Rathore, A.S., Bisen, D. et al. has done research on Performances of Machine Learning Models for Diagnosis of Alzheimer's Disease. The machine learning algorithms used in this paper are standards and successfully applied in classification problems. Along with classification algorithms, different feature selection and dimension reduction techniques are used for diffing out more relevant features than others for decision making and thus reducing the training time of the classification algorithms. In this study, Top-rated four features namely CDR, SES, nWBV, and EDUC are identified for decision making for AD that map sufficiently accurate correlation with the class labels and an approximately 90% accuracy.It has a low accuracy and only 3 classes of disease.

LIMITATIONS IN THE EXISTING SYSTEM

In the traditional way of approach Hospitals and clinic are the most widely used place by the patients to consult doctor and get treated, people consider it as the most reliable means to check their health status. But in this way of approach for treatment the patients need to wait for a long time to consult the doctor which makes them more sick .In the existing system the accuracy level is very less. In that system the disease recognised by the chatbot is not much accurate, because the disease is predicted only based upon the symptoms sent by the user, so that it may diagnose the disease wrongly. In those existing system there is only text feature and it is designed only for specific languages like French, norwegian and for specific domains. In the existing system the technique used was CNN , it has less number of layers than DNN so the prediction is less accurate.

PROPOSED SYSTEM

Input

Dataset is collected from the Kaggle. It contains MRI scan image of the patients having alzheimer. MRI scan images help us to clearly segment and study the image with fine detailing. The symptoms, risk factor and prevention of various diseases have been collected from the people for the chatbot.

Process

Initially, the libraries are imported and the dataset is loaded into the system. Dataset is then cleaned and

pre-processed. TensorFlow is used. It is an end-to-end open source platform for machine learning. Keras is also used as it is a high-level neural network library that runs on top of TensorFlow. Dataset is split into training and testing data. Epoch, an arbitrary cutoff, generally defined as "one pass over the entire dataset", used to separate training into distinct phases, which is useful for logging and periodic evaluation. When using validation data or validation split with the fit method of Keras models, evaluation will be run at the end of every epoch. Finally after scaling the data, a prediction model (mobilenet) is built and can be used to identify whether a person is affected by Alzheimer disease or not.

Output

The performance of the model is evaluated based on precision, accuracy rate. Confusion matrix for the test data is done. Confusion matrix is a performance measurement for machine learning classification problem where output can be of more classes. It is a table with 4 different combinations of rows and columns which is been predicted with actual values.

The patient can able to identify whether he/she has been affected by Alzheimer disease or not using this prediction model .The patient can able to analyze the symptoms , risk factor and prevention method of specific disease.

DESIGN DIAGRAM

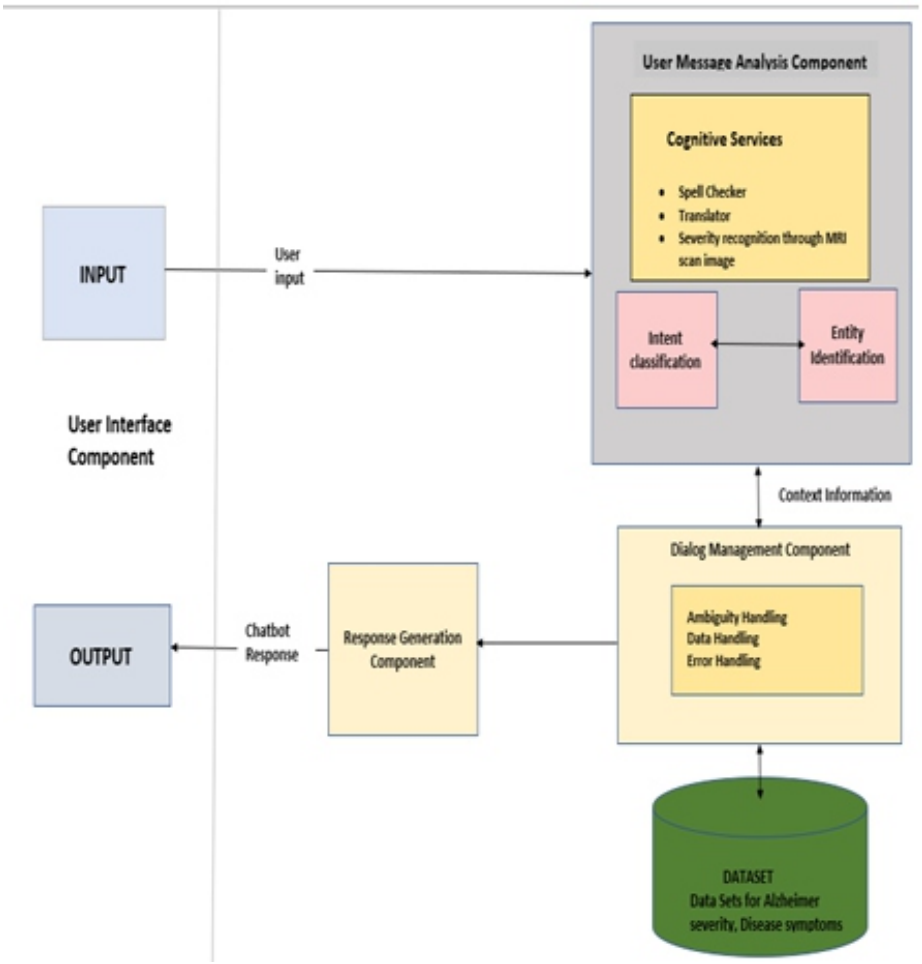

Figure 1. Detailed design diagram of a novel chatbot system

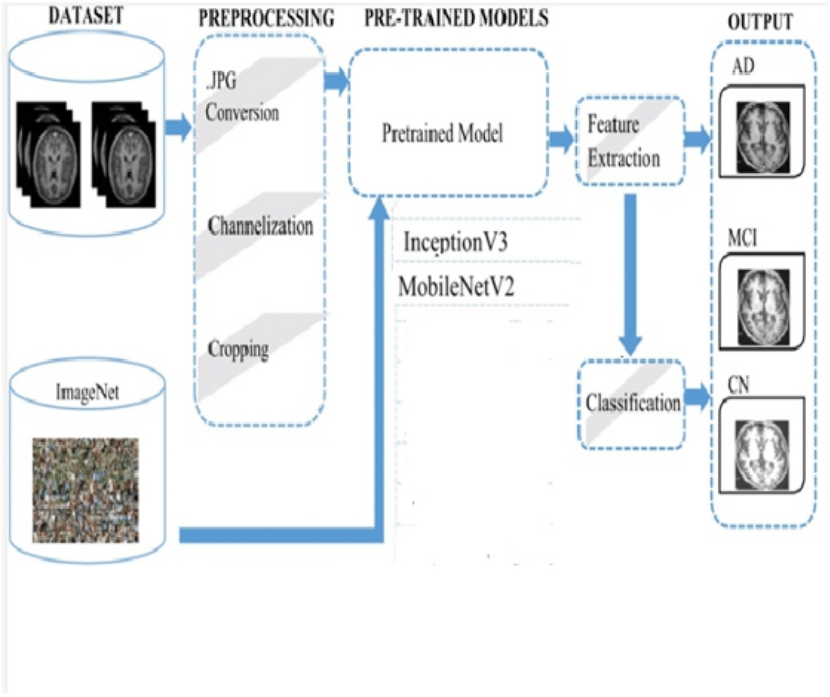

Figure 2. Process flow diagram for the alzheimer's disease detection

RESULT ANALYSIS

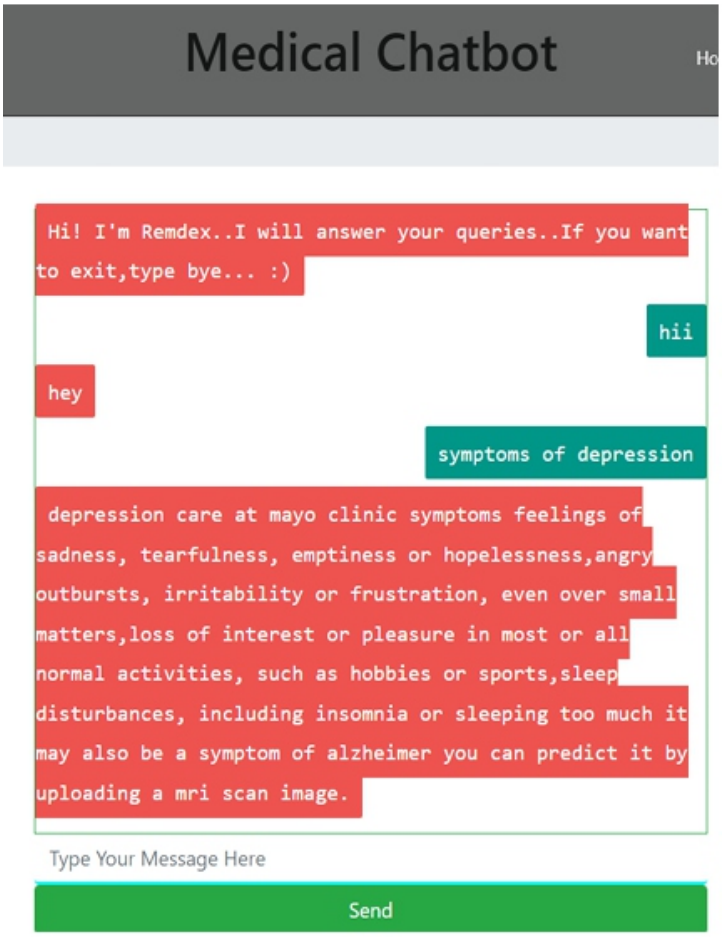

Figure 3. Chatbot for disease prediction

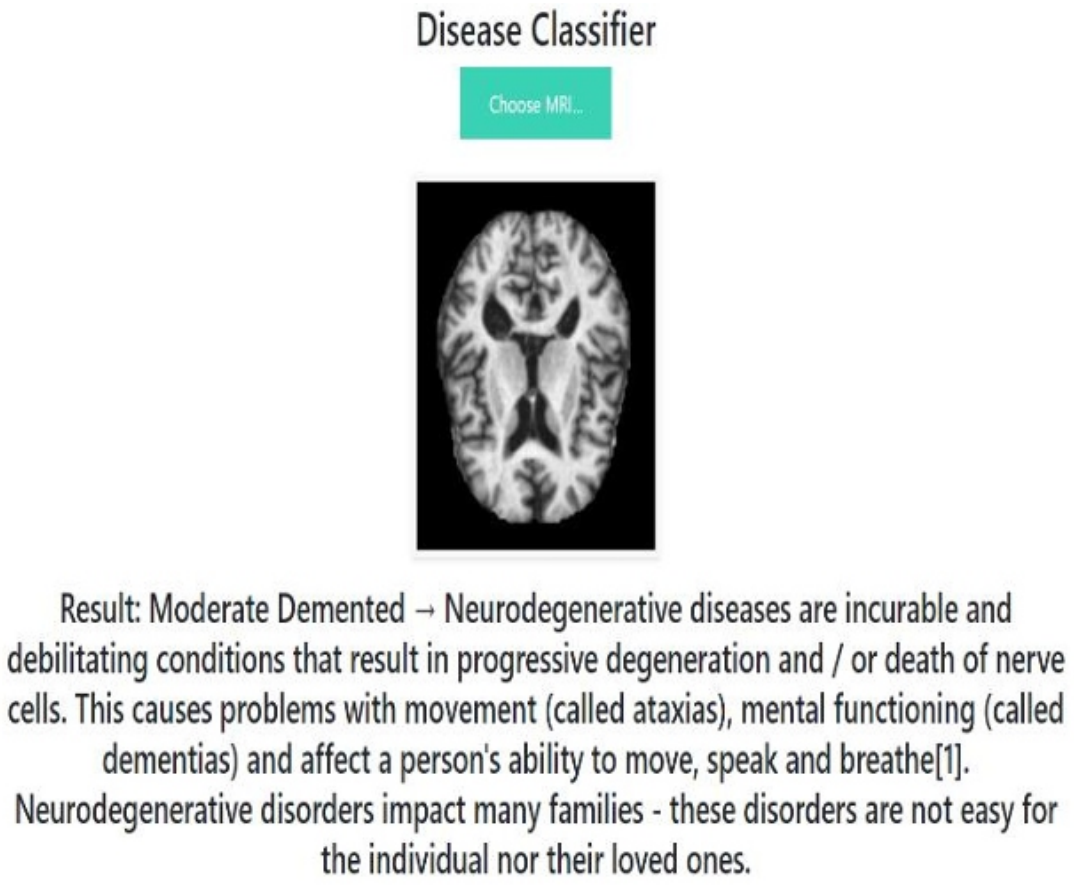

Figure 4. Alzheimer disease detection

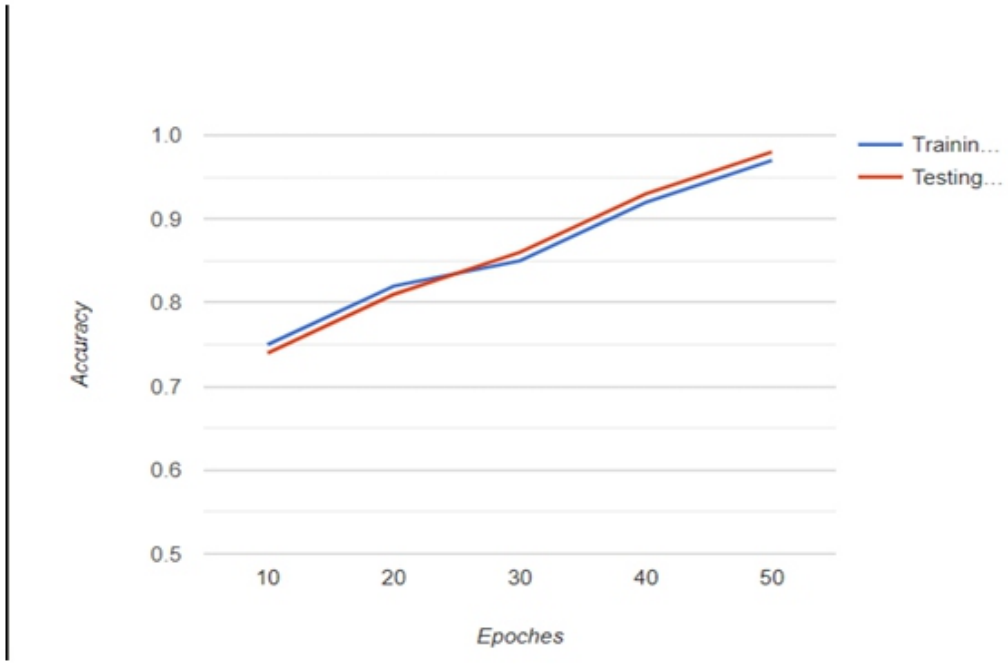

Figure 5. Accuracy graph for disease detection

CONCLUSION

In this system the people can able to diagnose the severity of disease easily from their respective locations. Here the application is developed to provide the response in a short period of time. In this system, it can able to function as a virtual doctor. It is highly difficult for working people to go to hospitals for their regular check-up. In such cases, this system is of great importance because it offers diagnostic support with a simple push of a button . The user interacts with the Prediction Engine by filling a form which holds the parameter set provided as an input to the trained models. This research has resulted in the development of a DNN-based pipeline to successfully identify multi-class Alzheimer's disease using brain MRI scan images and analyse the symptoms, risk factors and prevention of specific disease.

ACKNOWLEDGMENT

We are deeply indebted to Dr. P. Maragathavalli, Assistant Professor, Department of Information Technology, Puducherry Technological University, Puducherry, for her valuable guidance throughout the project work.

REFERENCES

[1] Uddin, M.Z., Dysthe, K.K., Følstad, A. et al. Deep learning for prediction of depressive symptoms in a large textual dataset. Neural Computing & Application 34,721–744 (2022). https://doi.org/10.1007/s00521-021-06426-4.

[2] Arafa, D.A., Moustafa, H.ED., Ali-Eldin, A.M.T. et al. Early detection of Alzheimer's disease based on the state-of-the-art deep learning approach: a comprehensive survey. Multimedia Tools Application 81, 23735–23776 https://doi.org/10.1016/j.cmpb.2021.106032. (2022).

[3] Shuangshuang Gao, Dimas Lima, A review of the application of deep learning in the detection of Alzheimer's disease, International Journal of Cognitive Computing in Engineering, Volume 3, 2022.

[4] Marwa EL-Geneedy, Hossam El-Din Moustafa, Fahmi Khalifa, Hatem Khater, Eman AbdElhalim, An MRI-based deep learning approach for accurate detection of Alzheimer's disease, Alexandria Engineering Journal, Volume 63, 2023, https://doi.org/10.1016/j.aej.2022.07.062.

[5] Arriba-Pérez, F., García-Méndez, S., González-Castaño, F.J. et al. Automatic detection of cognitive impairment in elderly people using an entertainment chatbot with Natural Language Processing capabilities. Journal of Ambient Intelligence Human Computing (2022).

[6] Pham, K.T., Nabizadeh, A. & Selek, S. Artificial Intelligence and Chatbots in Psychiatry. Psychiatry Q 93, 249–253 (2022) https://doi.org/10.1007/s11126-022-09973-8 LDV Forum - GLDV Journal for Computational Linguistics and Language Technology.

[7] Achtaich Khadija, Fagroud Fatima Zahra, Achtaich Naceur, AI-Powered Health Chatbots: Toward a general architecture, Procedia Computer Science, https://doi.org/10.1016/j.procs.2021.07.048. 2021,

[8] Dataset References: https://www.kaggle.com/datasets/itachi9604/disease-sympto m-description-dataset?select=dataset.csv https://www.kaggle.com/datasets/tourist55/alzheimers-datase t-4-class-of-images

Liveness Identity Verification for Face Anti-Spoofing in Biometric Validation using Recurrent Neural Network

P.Maragathavalli 1*, J.Sharmila 2, Syed Abdul Kareem 3, Nekkanti Bhavitha 4
1 Assistant Professor, Information Technology, Puducherry Technological University, Puducherry, India
2, 3, 4 B.Tech Student, Information Technology, Puducherry Technological University, Puducherry, India

A B S T R A C T

Face anti-spoofing is the task of preventing false facial verification by using a photo, video, mask or a different substitute for an authorized person's face. It has become an increasingly important and critical security feature for authentication systems, due to rampant and easily launchable presentation attacks. However, most previous approaches still suffer from diverse types of spoofing attacks, which are hardly covered by the limited number of training datasets, and thus they often show the poor accuracy when unseen samples are given for the test. To address this problem, a novel method is proposed based on liveness identity verification for face anti-spoofing in biometric validation using the Recurrent Neural Network (RNN).

Keywords : Biometric Validation, Face Anti-Spoofing Identification, Face Liveness Detection, Face Recognition, Lightweight CNN, Machine Learning, RNN.

INTRODUCTION

Faces can be captured conveniently by digital cameras, web cameras, smart phones, etc. The convenience is a double edged sword. It makes faces become not only the most widely used but also the most untrustable biometric modality. With the fast development of face recognition, the modern face recognition algorithms, especially deep networks trained on large scale datasets, can surpass human performance, but they may be easily fooled by face spoofing attacks which can be easily launched by inexperienced attackers. It is noteworthy that the proposed method only requires live facial images for training the model by using Recurrent Neural Network (RNN), which are easier to obtain than fake ones, and thus the generality power for resolving the problem of face anti-spoofing can be expected to be improved.

Experimental results on various benchmark datasets demonstrate the efficiency and robustness of the proposed method.

MOTIVATION

Face recognition on our mobile phones facilitates - Unlocking the device, Conducting financial transactions , Access to privileged content stored on the device. Failure to detect spoof attacks on smartphones could compromise confidential information such as emails, banking records, social media content, and personal photos. Biometric verification is a crucial activity in bank locker security system where the spoofing attack cannot be tolerated. Face-Anti Spoofing is used to minimize the fraudulent activities in the virtual interviews, online classes, online examinations where some unauthorized persons indulge in the activity of doing mal-practices.

LITERATURE SURVEY

S.No	NAME OF THE JOURNAL, YEAR	PAPER TITLE	JOURNAL DETAILS	TECHNIQUES USED	DEMERITS
1.	IEEE Transactions on Information Forensics and Security,2021	DRL-FAS: A Novel Framework Based on Deep Reinforcement Learning for Face Anti-Spoofing	R. Cai, H. Li, S. Wang, C. Chen and A. C. Kot, "DRL-FAS: A Novel Framework Based on Deep Reinforcement Learning for Face Anti-Spoofing," in IEEE Transactions on Information Forensics and Security, vol. 16, pp. 937 951,2021,doi:10.1109/TIF S.2020.3026553.	Face anti-spoofing, deep learning, reinforcement learning.	Live video identification is not done
2.	IEEE Journal ,2020	A Face Spoofing Detection Method Based on Domain Adaptation and Lossless Size Adaptation	W. Sun, Y. Song, H. Zhao and Z. Jin, "A Face Spoofing Detection Method Based on Domain Adaptation and Lossless Size Adaptation," in IEEE Access, vol. 8, pp. 66553-66563, 2020,doi:10.1109/ACCES S.2020.2985483.	Domain adaptation, face anti-spoofing, face liveness detection, face presentation attack detection, face spoofing detection, forensics, machine learning, pattern recognition.	This kind of attacks is less prevalent as compared to the photo attacks and the video attacks since it is relatively difficult to make a mask.3D mask attacks are usually not included in common face spoofing detection datasets.
3.	IEEE Journal ,2020	One-Class Learning Method Based on Live Correlation Loss for Face Anti-Spoofing	S. Lim, Y. Gwak, W. Kim, J. -H. Roh and S. Cho, "One-Class Learning Method Based on Live Correlation Loss for Face Anti-Spoofing,"in IEEE Access, vol. 8, pp. 201635-201648, 2020, doi: 10.1109/ACCESS.2020.30 35747.	Biometric authentication systems, face anti-spoofing, one-class learning, live correlation loss, feature correlation network.	Complexity level is high
4.	IEEE Transactions on Information Forensics and Security,2021	Camera Invariant Feature Learning for Generalized Face Anti-Spoofing	B. Chen, W. Yang, H. Li, S. Wang and S. Kwong, "Camera Invariant Feature Learning for Generalized Face Anti-Spoofing," in IEEE Transactions on Information Forensics and Security, vol. 16, pp. 2477-2492, 2021, doi: 10.1109/TIFS.2021.30550 18.	Face anti-spoofing, camera invariant, deep learning, generalization capability.	Image attacks are only identified and 3D mask attacks are usually not included in common face spoofing detection datasets.

Table 1. Literature Survey of the Existing Work

LIMITATIONS IN THE EXISTING SYSTEM

Generalizability

Since the exact type of spoof attack may not be known beforehand, how to generalize well to unknown 2D attacks is of utmost importance. A majority of the prevailing state-of-the-art face anti-spoofing techniques focus only on detecting 2D printed paper and video replay attacks, and are vulnerable to spoofs crafted from materials not seen.

Lack of Interpretability

Given a face image, face anti-spoofing approaches typically output a holistic face “spoons score” which depicts the likelihood that the input image is live or spoof. Without an ability to visualize which regions of the face contribute to the overall decision made be the network, the global spoons score alone is not sufficient for a human operator to interpret the network’s decision.

PROPOSED SYSTEM

Input

In this system, image acquisition is done and the 2C images is converted to 3D format. It helps us to

collect more details about the affected region in the image. Pickle Dataset is used in this Proposed system Liveness Identity Verification For Face Anti-Spoofing. Pickle in Python is primarily used in serializing and deserializing a Python object structure. It is the process of converting a Python object into a byte stream to store it in a file or database, maintain program state across sessions, or transport data over the network. At first Python pickle serialize the object and then converts the object into a character stream so that this character stream contains all the information necessary to reconstruct the object in another python script. Liveness Detection is carried out here by capturing the face of the person using integrated camera and with that dataset the various frames are generated and stored.

Process

A novel framework based on RNN for the FAS problem is proposed here. While many of the previous works used RNN to leverage temporal information from video frame. We use the advantage of RNN to memory information to reinforce extracted local features gradually. The collected data is then reduced by using 3D Discrete Cosine Transformation (DCT).Extraction of the facial characteristics is done by stemmer based feature extraction method. Feature extracted are reduced with the XGBoost Feature Reduction Technique by using the frames already generated which have the sequence of images of the person.

Output

The classifier that is used in the system is Recurrent Neural Network. The approach expected to provides the best-estimated accuracy of around 97% which identifies whether the person is authorized genuine person or the unauthorized spoof person.

FEATURES

Human Face is detected and characteristics are identified. Biometric validation of human traits with the Dataset. Stemmer Feature Extraction to reduce the dimensionality of the image.

XGBoost Feature Reduction for determining the results. 97% of accuracy in identification of spoofing attack.

DESIGN DIAGRAM

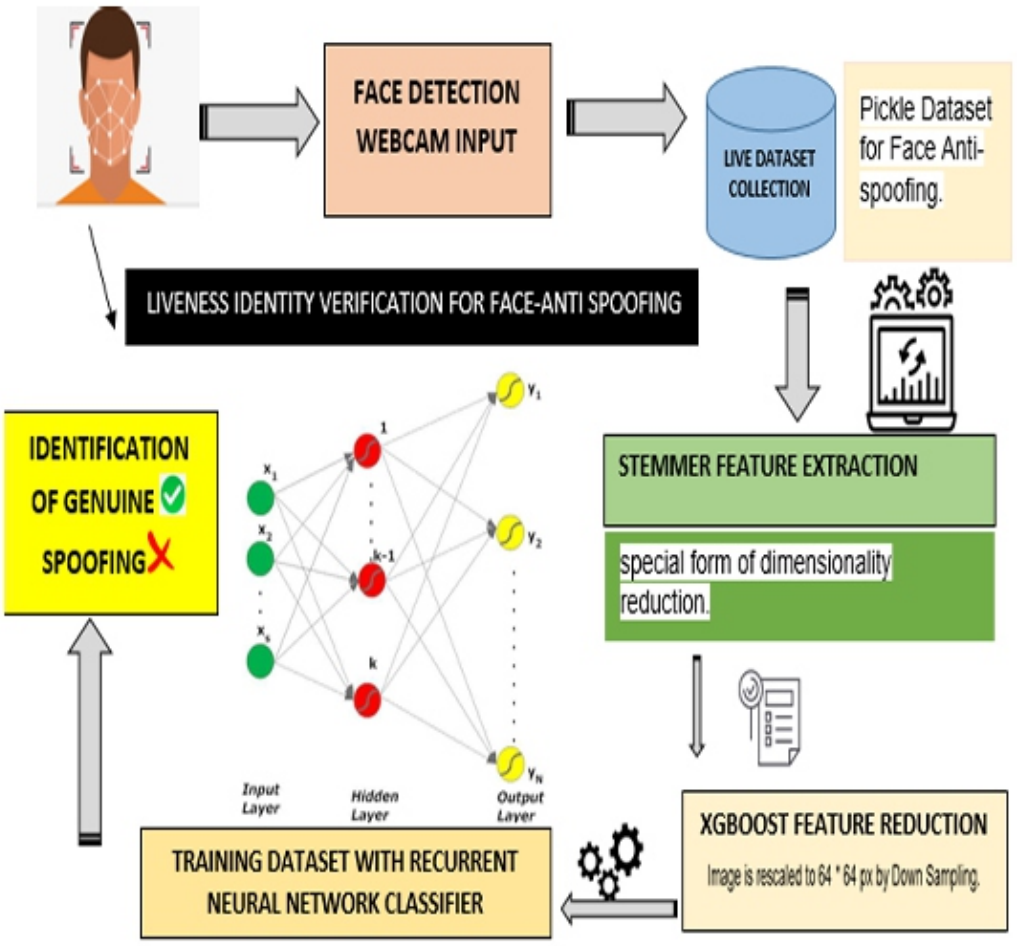

Figure 1. Detailed Design Diagram

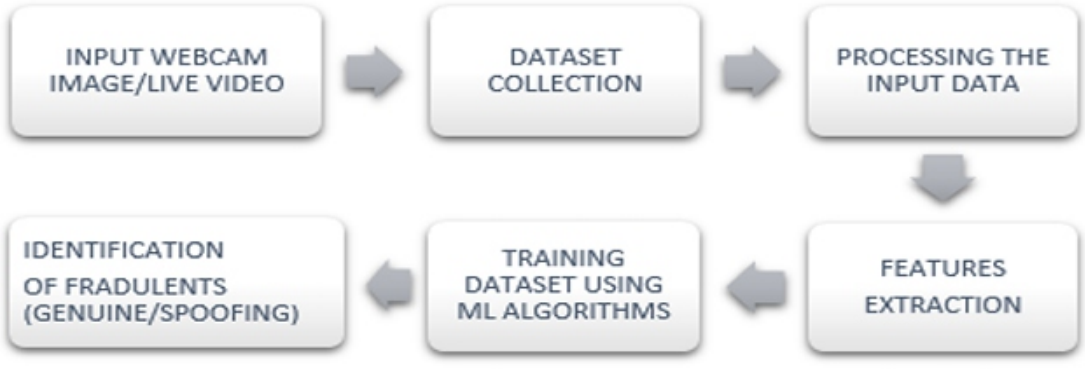

Figure 2. Process Flow Diagram

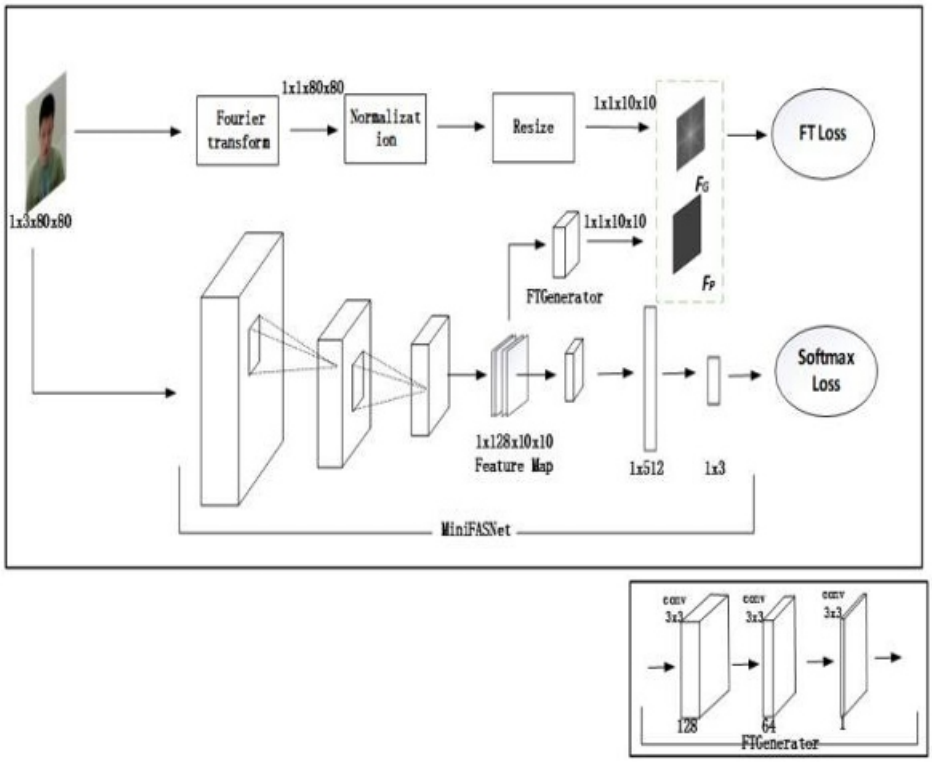

Figure 3. Recurrent Neural Network Architecture

RESULT ANALYSIS

Generating Frames

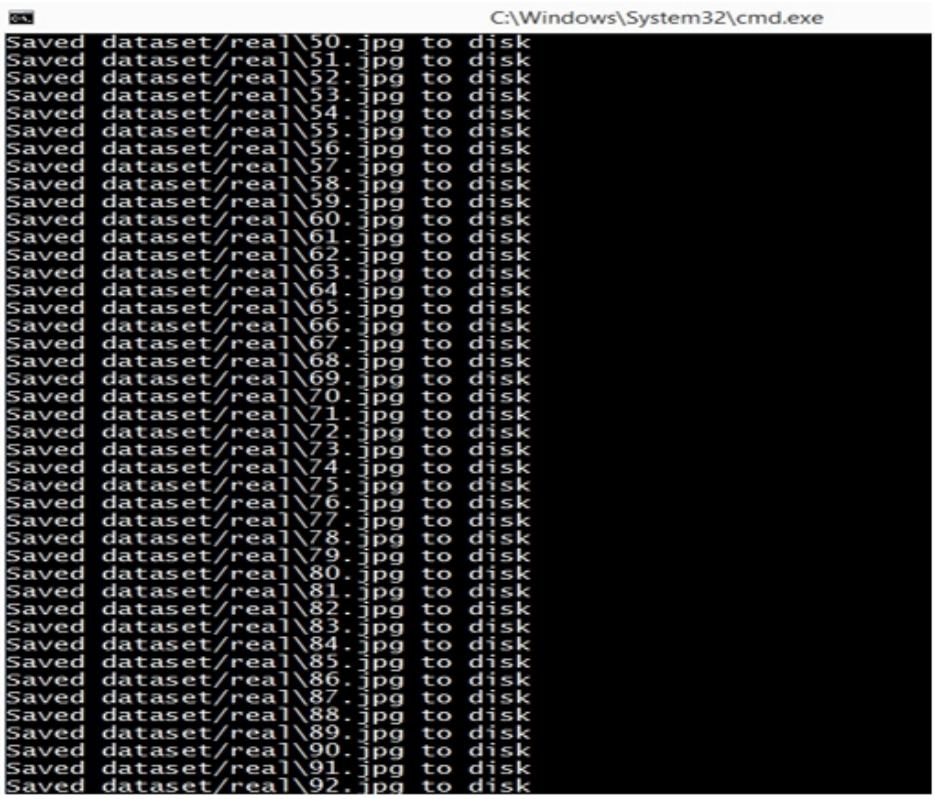

Figure 4. Frames Generation Results for Real and Fake Video Dataset

Training Model

```
Epoch 1/50
45/45 [==============================] - 9s 54ms/step - loss: 0.6127 - accuracy: 0.7430 - val_loss: 0.
7124 - val_accuracy: 0.3197
Epoch 2/50
45/45 [==============================] - 2s 34ms/step - loss: 0.2778 - accuracy: 0.9302 - val_loss: 0.
7036 - val_accuracy: 0.3197
Epoch 3/50
45/45 [==============================] - 2s 34ms/step - loss: 0.2758 - accuracy: 0.9167 - val_loss: 0.
6713 - val_accuracy: 0.9754
Epoch 4/50
45/45 [==============================] - 1s 29ms/step - loss: 0.2104 - accuracy: 0.9413 - val_loss: 0.
6509 - val_accuracy: 1.0000
Epoch 5/50
45/45 [==============================] - 1s 30ms/step - loss: 0.1708 - accuracy: 0.9777 - val_loss: 0.
5362 - val_accuracy: 1.0000
Epoch 6/50
45/45 [==============================] - 1s 31ms/step - loss: 0.1956 - accuracy: 0.9469 - val_loss: 0.
3906 - val_accuracy: 1.0000
Epoch 7/50
45/45 [==============================] - 1s 32ms/step - loss: 0.1654 - accuracy: 0.9609 - val_loss: 0.
2748 - val_accuracy: 1.0000
Epoch 8/50
45/45 [==============================] - 2s 38ms/step - loss: 0.1932 - accuracy: 0.9497 - val_loss: 0.
2031 - val_accuracy: 1.0000
Epoch 9/50
45/45 [==============================] - 3s 56ms/step - loss: 0.1305 - accuracy: 0.9721 - val_loss: 0.
1257 - val_accuracy: 1.0000
```

Figure 5. Training Results of FAS Identification Model

Training Loss & Accuracy on Dataset

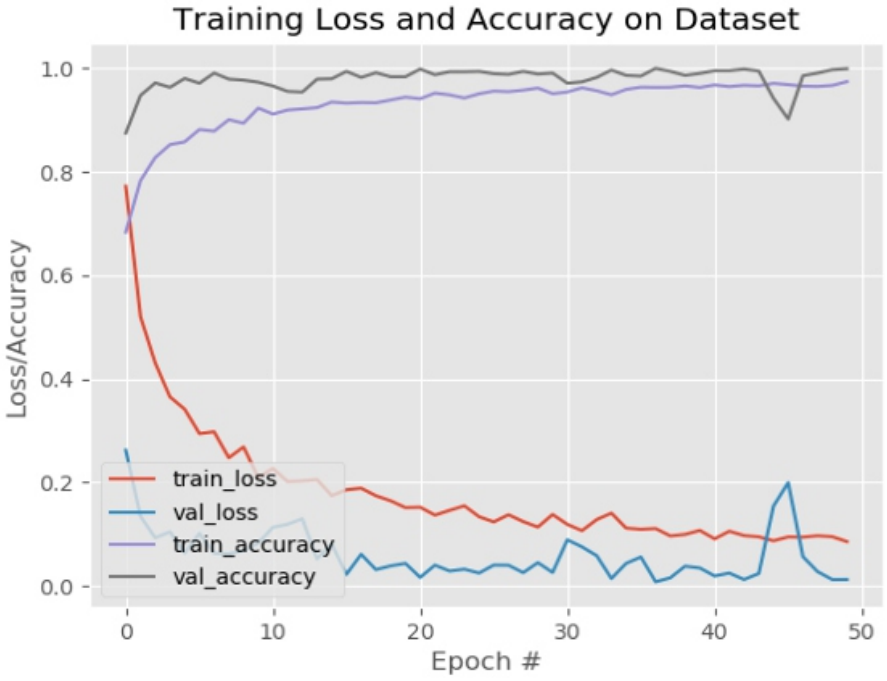

Figure 6. Validation and Testing Accuracy of FAS Model

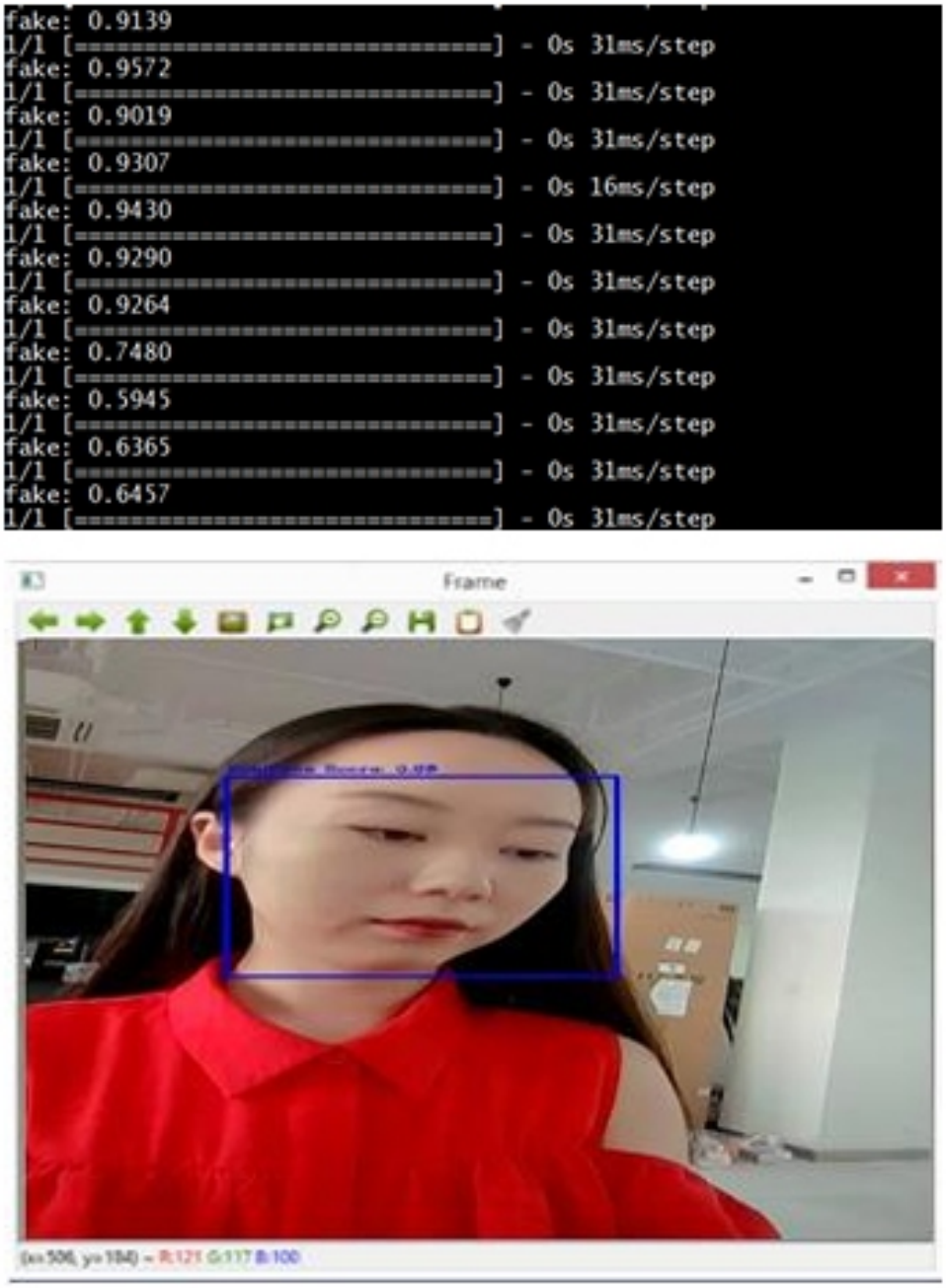

Figure 7. Result for Identification of Real/Genuine Face

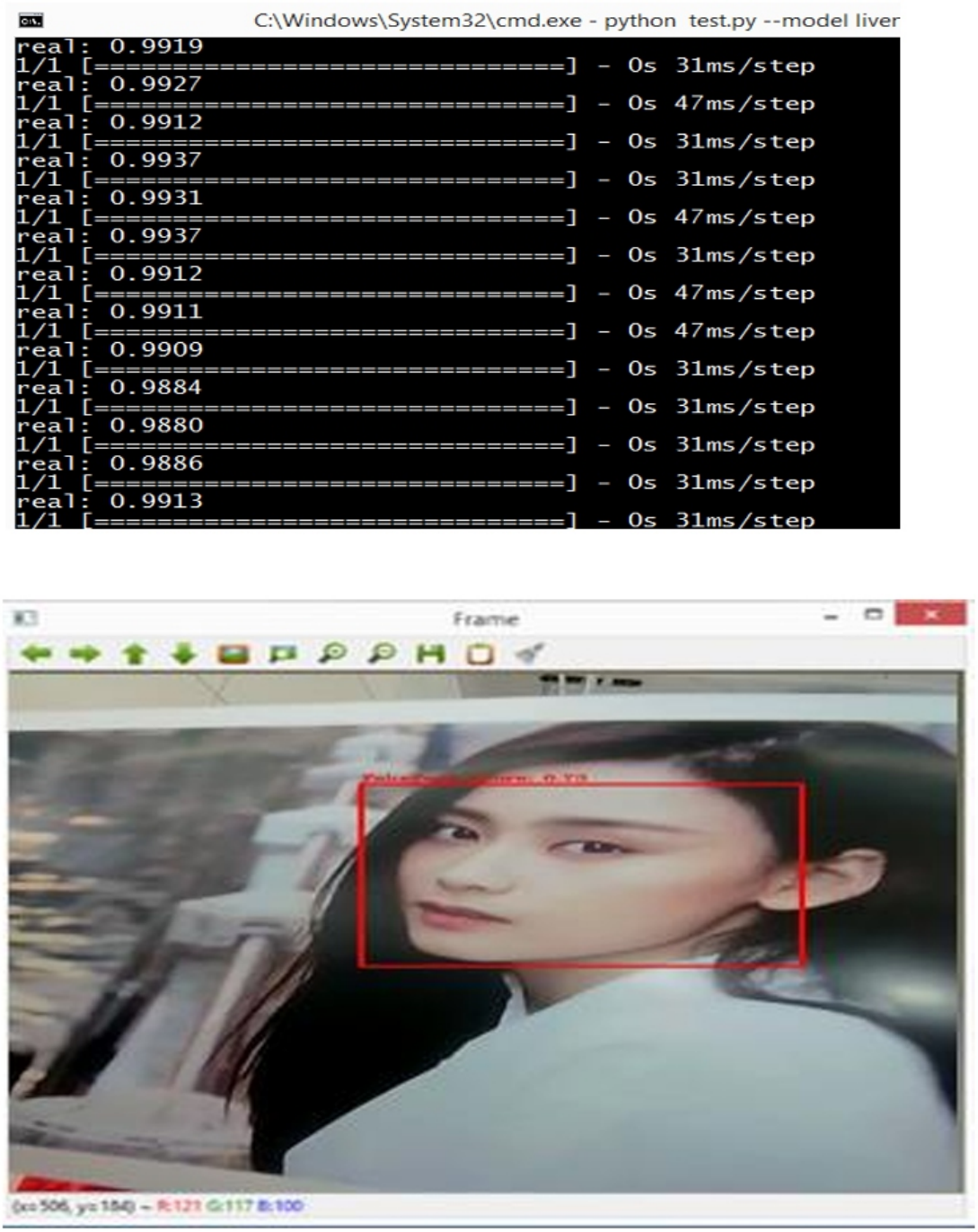

Figure 8. Result for Identification of Fake/Spoof Face

CONCLUSION

Under this face recognition approach, a user is required to take a special action called a challenge for liveness detection. The system ensures that required action was taken. Usually, a group of actions is required to make the model reliable. These actions can include smiles, expressions of emotions such as sadness, surprise or head movements. These interactions require significant time and are inconvenient for users. Face presentation attack detection is often considered as a binary classification task which results in over-fitting to the known attacks leading to poor generalization against unseen attacks. This system employs strengthen techniques enhance reputation price and execution time by using 3D DCT

Descriptor for face recognition and stemmer feature extraction and XGBoost feature reduction techniques with the help of Recurrent Neural Network classifier to detect the person indulge in the fraudulent activities. The experimental results show that the proposed anti-spoofing framework can prevent diversity of face attacking forms, such as dim light, realistic face camouflage, static or motion pattern in the most effective way. This system serves for various domains such as it helps to identify the spoofing attack in bank locker security system, Virtual Interviews, Online classes, Online examinations and in the highly-secured authentication systems.

ACKNOWLEDGMENT

We are deeply indebted to Dr. P. Maragathavalli, Assistant Professor, Department of Information Technology, Puducherry Technological University, Puducherry, for her valuable guidance throughout the project work.

REFERENCES

[1] R. Cai, H. Li, S. Wang, C. Chen and A. C. Kot, "DRL-FAS: A Novel Framework Based on Deep Reinforcement Learning for Face Anti-Spoofing," in IEEE Transactions on Information Forensics and Security, vol. 16, pp. 937-951, 2021, doi: 10.1109/TIFS.2020.3026553.

[2] A Face Spoofing Detection Method Based on Domain Adaptation and Lossless Size Adaptation W. Sun, Y. Song, H. Zhao and Z. Jin, "A Face Spoofing Detection Method Based on Domain Adaptation and Lossless Size Adaptation," in IEEE Access, vol. 8, pp. 66553-66563, 2020, doi: 10.1109/ACCESS.2020.2985453.

[3] One-Class Learning Method Based on Live Correlation Loss for Face Anti-Spoofing S. Lim, Y. Gwak, W. Kim, J. -H. Roh and S. Cho, "One-Class Learning Method Based on Live Correlation Loss for Face Anti-Spoofing,"in IEEE Access, vol. 8, pp. 201635-201648, 10.1109/ACCESS.2020.3035747. 2020, doi:

[4] B. Chen, W. Yang, H. Li, S. Wang and S. Kwong, "Camera Invariant Feature Learning for Generalized Face Anti-Spoofing," in IEEE Transactions on Information Forensics and Security, vol. 16, pp. 2477-2492, 2021, doi: 10.1109/TIFS.2021.3055018.

[5] D. Deb and A. K. Jain, "Look Locally Infer Globally: A Generalizable Face Anti-Spoofing Approach," in IEEE Transactions on Information Forensics and Security, vol. 16, pp. 1143-1157, 2021, doi: 10.1109/TIFS.2020.3029879.

[6] A. George and S. Marcel, "Learning One Class Representations for Face Presentation Attack Detection Using Multi-Channel Convolutional Neural Networks," in IEEE Transactions on Information Forensics and Security, vol. 16, pp. 361-375, 2021, doi: 10.1109/TIFS.2020.3013214.

[7] Data-Fusion-Based Two-Stage Cascade Framework for Multimodality Face Anti-Spoofing W. Liu, X. Wei, T. Lei, X. Wang, H. Meng and A. K. Nandi, "Data-Fusion-Based

Two-Stage Cascade Framework for Multimodality Face Anti-Spoofing," in IEEE Transactions on Cognitive and Developmental Systems, vol. 14, no. 2, pp. 672-683, June 2022, doi:10.1109/TCDS.2021.3064679.

[8] Datasets: https://www.kaggle.com/code/rspadim/reading-pickle-files https://www.kaggle.com/datasets?search=face+anti+spoofing

Instructions for Authors

Essentials for Publishing in this Journal

1 Submitted articles should not have been previously published or be currently under consideration for publication elsewhere.

2 Conference papers may only be submitted if the paper has been completely re-written (taken to mean more than 50%) and the author has cleared any necessary permission with the copyright owner if it has been previously copyrighted.

3 All our articles are refereed through a double-blind process.

4 All authors must declare they have read and agreed to the content of the submitted article and must sign a declaration correspond to the originality of the article.

Submission Process

All articles for this journal must be submitted using our online submissions system. http://enrichedpub.com/ . Please use the Submit Your Article link in the Author Service area.

Manuscript Guidelines

The instructions to authors about the article preparation for publication in the Manuscripts are submitted online, through the e-Ur (Electronic editing) system, developed by **Enriched Publications Pvt. Ltd**. The article should contain the abstract with keywords, introduction, body, conclusion, references and the summary in English language (without heading and subheading enumeration). The article length should not exceed 16 pages of A4 paper format.

Title

The title should be informative. It is in both Journal's and author's best interest to use terms suitable. For indexing and word search. If there are no such terms in the title, the author is strongly advised to add a subtitle. The title should be given in English as well. The titles precede the abstract and the summary in an appropriate language.

Letterhead Title

The letterhead title is given at a top of each page for easier identification of article copies in an Electronic form in particular. It contains the author's surname and first name initial .article title, journal title and collation (year, volume, and issue, first and last page). The journal and article titles can be given in a shortened form.

Author's Name

Full name(s) of author(s) should be used. It is advisable to give the middle initial. Names are given in their original form.

Contact Details

The postal address or the e-mail address of the author (usually of the first one if there are more Authors) is given in the footnote at the bottom of the first page.

Type of Articles

Classification of articles is a duty of the editorial staff and is of special importance. Referees and the members of the editorial staff, or section editors, can propose a category, but the editor-in-chief has the sole responsibility for their classification. Journal articles are classified as follows:

Scientific articles:

1. Original scientific paper (giving the previously unpublished results of the author's own research based on management methods).

2. Survey paper (giving an original, detailed and critical view of a research problem or an area to which the author has made a contribution visible through his self-citation);

3. Short or preliminary communication (original management paper of full format but of a smaller extent or of a preliminary character);

4. Scientific critique or forum (discussion on a particular scientific topic, based exclusively on management argumentation) and commentaries. Exceptionally, in particular areas, a scientific paper in the Journal can be in a form of a monograph or a critical edition of scientific data (historical, archival, lexicographic, bibliographic, data survey, etc.) which were unknown or hardly accessible for scientific research.

Professional articles:

1. Professional paper (contribution offering experience useful for improvement of professional practice but not necessarily based on scientific methods);
2. Informative contribution (editorial, commentary, etc.);
3. Review (of a book, software, case study, scientific event, etc.)

Language

The article should be in English. The grammar and style of the article should be of good quality. The systematized text should be without abbreviations (except standard ones). All measurements must be in SI units. The sequence of formulae is denoted in Arabic numerals in parentheses on the right-hand side.

Abstract and Summary

An abstract is a concise informative presentation of the article content for fast and accurate Evaluation of its relevance. It is both in the Editorial Office's and the author's best interest for an abstract to contain terms often used for indexing and article search. The abstract describes the purpose of the study and the methods, outlines the findings and state the conclusions. A 100- to 250-Word abstract should be placed between the title and the keywords with the body text to follow. Besides an abstract are advised to have a summary in English, at the end of the article, after the Reference list. The summary should be structured and long up to 1/10 of the article length (it is more extensive than the abstract).

Keywords

Keywords are terms or phrases showing adequately the article content for indexing and search purposes. They should be allocated heaving in mind widely accepted international sources (index, dictionary or thesaurus), such as the Web of Science keyword list for science in general. The higher their usage frequency is the better. Up to 10 keywords immediately follow the abstract and the summary, in respective languages.

Acknowledgements

The name and the number of the project or programmed within which the article was realized is given in a separate note at the bottom of the first page together with the name of the institution which financially supported the project or programmed.

Tables and Illustrations

All the captions should be in the original language as well as in English, together with the texts in illustrations if possible. Tables are typed in the same style as the text and are denoted by numerals at the top. Photographs and drawings, placed appropriately in the text, should be clear, precise and suitable for reproduction. Drawings should be created in Word or Corel.

Citation in the Text

Citation in the text must be uniform. When citing references in the text, use the reference number set in square brackets from the Reference list at the end of the article.

Footnotes

Footnotes are given at the bottom of the page with the text they refer to. They can contain less relevant details, additional explanations or used sources (e.g. scientific material, manuals). They cannot replace the cited literature.
The article should be accompanied with a cover letter with the information about the author(s): surname, middle initial, first name, and citizen personal number, rank, title, e-mail address, and affiliation address, home address including municipality, phone number in the office and at home (or a mobile phone number). The cover letter should state the type of the article and tell which illustrations are original and which are not.

Note