ENSEMBLE MODEL WITH HYBRID FEATURES FOR VOICE SPOOF DETECTION

Medikonda Neelima[1], I Santiprabha[2]
[1]Research Scholar, Department of Electronics and Communication Engineering, University College of Engineering, JNTUK, Kakinada, India.
[2]Professor, Department of Electronics and Communication Engineering, University College of Engineering, JNTUK, Kakinada, India.

***Abstract*—**The simplicity of voice communication has led to the emergence of speech-based biometric authentication systems. However, these systems are vulnerable to spoofing attacks where individuals attempt to exploit security credentials. This poses a significant threat to the reliability of automatic speaker verification (ASV) authentication systems. To address this issue, the research paper proposes the use of deep learning (DL) methods and ensembles of neural networks. The study outlines two different models: one with time-distributed dense, long short-term memory (LSTM) layers, and another based on temporal convolution (TC). An ensemble model incorporating these DNNs is also analyzed. The model utilizes hybrid features which combine the Mel frequency cepstral coefficients (MFCC), Inverted Mel Frequency Cepstral Coefficients (IMFCC), and Local Binary Pattern (LBP) features. The model is evaluated using the impersonation speech samples and ASVspoof2019 dataset which contains logical access (LA) and physical access (PA) spoofing attacks. The proposed solution aims to improve spoof detection and create systems capable of handling unknown data during testing. The results show promising outcomes, paving the way for further research in this domain using DL.

Keywords*—*deep learning, LSTM, audio spoofing, CNN

I. INTRODUCTION

Voice spoofing involves using technology, such as voice morphing software or deep learning algorithms, to impersonate someone else's voice. This technique can deceive individuals, organizations, or voice recognition systems used for security or authentication purposes. Voice spoofing attacks can lead to serious consequences, including monetary extortion, wholesale fraud, and unapproved admittance to delicate data.

The risk of voice spoofing has led to increased research and development in voice spoof detection. This process involves identifying and differentiating genuine voices from spoofed voices using various techniques, such as analyzing spectral features, detecting artifacts introduced by the spoofing technique, or comparing the voice with a known reference voice. However, many techniques struggle with detecting unknown attacks that have different statistical distributions from known attacks, which can pose practical difficulties [10].

Our main objective is to develop a security layer that safeguards customers from voice replay, voice conversion, speech synthesis, and impersonation spoofing attacks on the ASV systems [2]. Reliable voice spoof detection systems are vital for ensuring the security and integrity of voice-based applications and services. With the increasing utilization of voice-based technologies, the demand for dependable voice spoof detection solutions will continue to escalate. The proposed work is given in figure 1.

Fig. 1. Overview of proposed work

II. LITERATURE SURVEY

Voice spoofing detection using LSTM models is a relatively new research area, but there have been several studies and papers published on this topic. Here is a brief literature review of some of the key research in this area:

"Voice Spoofing Detection Using Recurrent Neural Networks" by Han et al. (2017): This study explored the use of recurrent neural networks (RNNs), including LSTM models, for detecting voice spoofing attacks. The authors used a dataset of genuine and spoofed audio recordings and achieved high accuracy in detecting spoofed audio using their RNN model.

"Voice Spoofing Detection Using Spectral Features and Long Short-Term Memory Networks" by Ming and Li (2018): This study proposed a voice spoofing detection system that combined spectral features with LSTM models.

"Deep Learning for Voice Spoofing Detection: A Comprehensive Review" by Al-Taweel and Al-Dhabyani (2020): This paper provided a comprehensive review of the use of deep learning techniques, including LSTM models, for voice spoofing detection.

"Voice Spoofing Detection Using Convolutional Neural Networks and Long Short-Term Memory" by Rahaman et al. (2021): This study proposed a voice spoofing detection system that combined convolutional neural networks (CNNs) and LSTM models.

"Voice spoofing countermeasure for voice replay attacks using deep learning," by Zhou et al. (2022): This study proposed work on the BiLSTM model to achieve better performance and reliability of the model.

Generally speaking, these investigations recommend that LSTM models can be a viable instrument for voice spoofing detection, particularly when combined with other deep learning techniques or spectral features. However, more exploration is expected to investigate the robustness and scalability of these models in real-world scenarios. In the previous research, it is understood that even though many models are proposed for voice spoof detection but they are not completely effective in detecting voice spoofs.

III. DATASET

ASVspoof 2019 dataset was released by the ASV community in 2019. the dataset was created to support research on the development of countermeasures for automatic speaker verification systems, which are vulnerable to spoofing attacks. The database contains speech recordings of both genuine and spoofed utterances, which were generated using various techniques such as replay, voice conversion, and speech synthesis. The impersonation voice samples are collected from celebrity's speeches [11]. Data samples used in the proposed work are given in table I.

TABLE I. OVERALL DATASET

S. No	*Type Of Samples*	*Total Samples*	*Used Samples*
1	Replay Genuine	5400	1000
2	Replay Spoof	16444	1000
3	Logical Access Genuine	2700	1000
4	Speech Synthesis Spoof	16444	1000
5	Voice Conversion Spoof	7500	1000
6	Impersonation Genuine	50	50
7	Impersonation Spoof	55	50

IV. FEATURE EXTRACTION

Feature extraction is a crucial step in any speech recognition system as it enables the identification of linguistic content and the removal of unwanted information such as noise and emotions. This project employs three types of feature extraction methods: MFCC, IMFCC, and LBP.

A. Mel-frequency Cepstral Coefficients (MFCC)

Throughout the long term, for speech-related applications, when the human auditory system has become a standard acoustic feature set, the Mel-Frequency Cepstral Coefficients (MFCC) were modeled. Recently, authors have demonstrated that the Inverted Mel Frequency Cepstral Coefficients (IMFCC) give a helpful list of features for speaker identification by containing corresponding data present in high-frequency regions [3].

The MFCC feature extraction technique is frequently used in voice spoof detection to differentiate between genuine and fake audio recordings. This method captures the spectral envelope of a sound signal and converts it into a set of coefficients that represent the spectral shape of the signal. MFCC feature extraction method is effective in distinguishing between genuine and fake audio recordings.

This approach to feature extraction holds promise and may improve results [5]. Overall, MFCC feature extraction is a powerful technique in voice spoof detection and has demonstrated high accuracy in distinguishing between genuine and fake audio recordings.

B. Improved Mel-frequency Cepstral Coefficients (IMFCC)

IMFCC (Improved Mel-frequency Cepstral Coefficients) is a feature extraction method utilized in voice spoof detection to analyze and distinguish between genuine and fake voices [8]. It is an improved version of MFCCs, which are ordinarily utilized in speech recognition and signal processing applications. The IMFCC is used to catch speaker explicit data lying in the higher frequency part of the spectrum and is normally disregarded by MFCC [3].

IMFCCs are based on the concept of the Mel scale, which is a perceptual size of pitches that is based on the way the human ear perceives sounds. The Mel scale is divided into a series of overlapping frequency bands that are logarithmically spaced. In the first step of the IMFCC feature extraction process, the audio signal is divided into short segments, and for each segment, the power spectrum is calculated using the Fourier transform.

Next, the power spectrum is passed through a filter bank that mimics the Mel scale, which helps to emphasize the frequency bands that are more relevant to human perception of sound. The resulting filter bank output is then transformed using the discrete cosine transform (DCT), which generates a set of coefficients known as cepstral coefficients.

Finally, the IMFCC method improves upon the MFCC method by applying a logarithmic compression and a normalization step to the cepstral coefficients. This helps to reduce the effects of noise and variability in the audio signal, making the feature extraction process more robust.

In voice spoof detection, IMFCCs can be used as input to machine learning algorithms that are trained to distinguish between genuine and fake voices. By analyzing the differences in IMFCC features between genuine and fake voices, these algorithms can accurately detect and prevent voice spoofing attacks.

The main difference between IMFCC and MFCC is that IMFCC uses a different weighting scheme to calculate the Mel filter bank coefficients. Instead of using triangular filters, IMFCC uses a raised-cosine filter bank, which provides a better resolution in the frequency domain.

C. Local Binary Pattern (LBP) Feature Extraction

Local Binary Pattern (LBP) feature extraction technique is carried out by the following steps.

Pre-processing: Initially, the raw audio signal undergoes pre-processing to eliminate any noise or artifacts that could disrupt the feature extraction process. This may involve filtering, normalization, and other methods.

Frame segmentation: After pre-processing, the signal is segmented into brief frames, usually lasting between 20-30 ms. Each frame undergoes independent analysis.

Feature extraction: A 2D array of samples is created for each frame, where the rows relate to varying frequency bands and the columns correspond to distinct time instants.

LBP operator: To generate a binary code for each sample, an LBP operator is utilized on the 2D array. This operator determines the difference between the central sample and its eight neighbors and assigns a binary value (0 or 1) to each neighbour based on whether its amplitude exceeds or falls below that of the central sample.

Histogram: The number of times each binary code appears in the 2D array is counted to create a histogram. This histogram indicates the distribution of local spatial patterns within the audio signal.

Normalization: To mitigate the influence of signal variations, like alterations in amplitude or background noise, the histogram is normalized.

Classification: Next, the LBP features are inputted to a machine learning algorithm, such as a neural network, that undergoes training to categorize the input into genuine or fake. Based on the LBP features, the algorithm predicts and outputs a result.

To differentiate between genuine and fake voices in voice spoof detection, the LBP feature extraction procedure involves numerous steps for analyzing the audio signal and extracting relevant features.

V. PROPOSED MODEL

The work introduced in this paper attempts to resolve the issue by utilizing deep learning (DL) techniques and an ensemble of various neural networks [1] [4] as shown in figure 1. Various blocks of the proposed model are discussed.

A. Long Short Term Memory (LSTM) unit

Long Short-Term Memory (LSTM) is a kind of Recurrent Neural Network (RNN) that can overcome the issue of gradient descent vanishing or exploding found in RNNs. RNNs have an inner memory that assists them with monitoring past data handled within the network, which can be utilized for detecting patterns. However, the drawback of vanishing/exploding gradient descents impairs their performance. LSTM is the extension of RNN that addresses this problem and allows the network to maintain long-term memory. LSTM units have gates such as Forget, Input, and Output gates that control the stream of data to and fro of the cell. LSTM has shown superior performance compared to RNNs in the classification, prediction, and processing of time-series data. It stores past input data in the memory cell while the gates regulate the information flow. The effectiveness of LSTM makes it suitable for solving problems related to NLP, and speech processing. LSTM can be leveraged in analyzing the acoustic features of voice signals to determine if the voice is genuine or spoofed. LSTM is well-suited for analyzing time-series data, including audio signals because it can master short as well as long-term sequential data. LSTM employs gates that enhance its capacity to capture non-linear association and feedback attachments, thus allowing it to better interpret the patterns of the data. This ability enables LSTM to be effective in voice spoof detection [7]. The work flow of LSTM unit is given in figure 2.

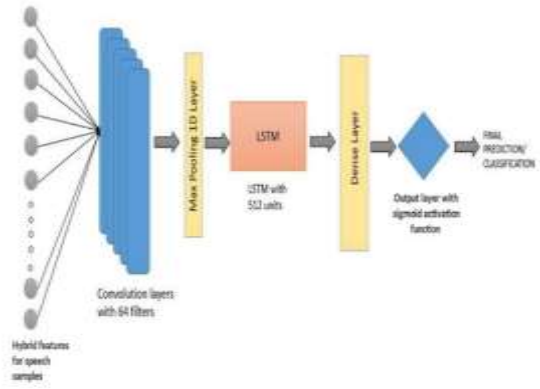

Fig. 2. LSTM model

B. Convolutional Neural Network (CNN)

CNNs, or Convolutional Neural Networks, are artificial neural networks that have been widely utilized in image processing and recognition. In a convolutional layer, a filter is used to apply convolution on the input image to extract features. Similarly, the extracted feature vectors from a speech signal are applied as images into a CNN.

A Temporal Convolutional Network (TCN) is a form of a neural network comprising convolutional layers that can process temporal data. CNN has proven to be an effective deep learning algorithm for voice spoof detection by training on a dataset of genuine and spoofed voice samples to learn the patterns that differentiate them, which are applied using a set of filters on the input data.

The filters utilized in CNNs are capable of recognizing essential patterns in audio relevant to voice spoof detection, such as the presence of specific frequency components or spectral patterns. After training, a CNN can be employed with high accuracy to classify new audio samples as genuine or spoofed. Various studies on voice spoof detection have demonstrated that this approach outperforms traditional machine learning methods based on handcrafted features. Investigations on the ASVspoof 2019 database indicate that CNNs yield systems comparable to or surpass state-of-the-art approaches for both physical and logical access attacks [6]. In general, CNNs provide an effective and powerful technique for detecting voice spoofing in real-world applications.

C. Temporal Convolutional Network (TCN)

TCN (Temporal Convolutional Network) is a deep learning algorithm that has shown significant potential for voice spoof detection. By analyzing raw audio signals, TCN can identify patterns that distinguish genuine speech from spoofed speech. The network excels at detecting long-term dependencies in the

input signal [9], which is crucial in voice spoof detection because spoofed speech may exhibit temporal patterns not present in unaltered speech.

TCN uses dilated convolution operations to extract features from input signals across various scales, which enables the network to capture long-term temporal dependencies efficiently. The output of each convolutional layer is passed through an activation function, which introduces non-linearity into the network and allows it to learn more complex patterns

Furthermore, TCN comprises residual connections that enhance gradient flow through the network and alleviate the vanishing gradient problem. These connections enable the network to learn deeper representations of input data, which can improve overall performance.

TCN provides a robust and efficient approach for detecting voice spoofing in practical scenarios. Its capacity to manage long-term temporal dependencies in input signals makes it particularly suitable for identifying complex patterns present in spoofed speech. The work flow of TCN is explained as follows.

Input layer: It receives the raw audio data in TCN. The input data is usually pre-processed by normalizing the sample rate and amplitude before being fed into the network.

Temporal network: The temporal convolutional layers are the central component of TCN architecture. They employ dilated convolution operations to extract features from the audio signal at various scales. This operation enables the network to capture long-term temporal dependencies within the audio signal, which is crucial in voice spoof detection.

Activation functions: Following each convolutional layer, an activation function is employed to the output. This introduces non-linearity into the network, which enables it to learn more complex relationships between input data and the target output.

Residual connections: Residual connections are utilized to enhance gradient flow through the network and address the vanishing gradient problem. They enable the network to learn deeper representations of input data.

Fully connected layers: The fully connected layers receive the output of the convolutional and residual layers and map it to the final classification outcome. A SoftMax activation function typically follows these layers to generate the probability of the input audio sample being genuine or spoofed.

Output layer: The output layer generates the final classification outcome by considering the probabilities generated by the fully connected layers.

VI. RESULTS AND DISCUSSION

Table II shows the f1-score which is (1.00,0.84), for Spoof and genuine respectively. The precision is (1.00,0.80) for spoof and genuine respectively. The recall is (1.00,0.89) for spoof and genuine respectively. The performance metrics also show better performance in training loss as well as validation loss which is shown in figure 3.

TABLE II. PERFORMANCE METRICS FOR THE PROPOSED SYSTEM

Speech Utterance type	*Precision*	*Recall*	*F1-score*
Spoof	1.00	1.00	1.00
Genuine	0.80	0.89	0.84

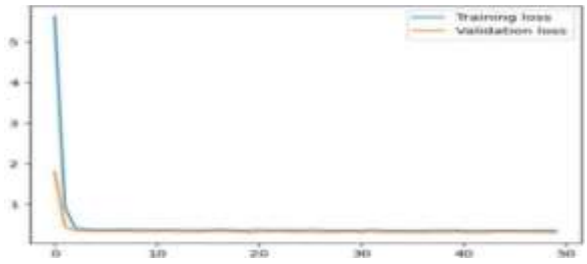

Fig. 3. Graph showing Training loss and Validation loss

The results were shown by the proposed model that detects a given genuine input sample as a genuine sample which is given in figure 4.

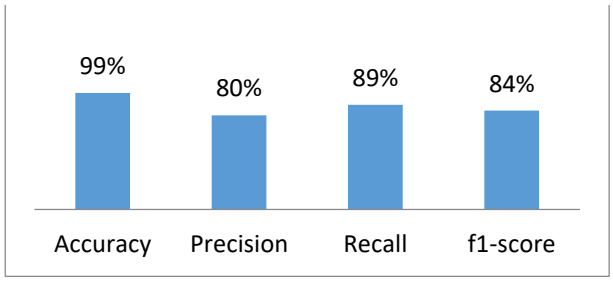

Fig. 4. Results for detecting genuine samples

For different spoof samples, the detection rate obtained by the proposed model is very high which is shown in figure 5.

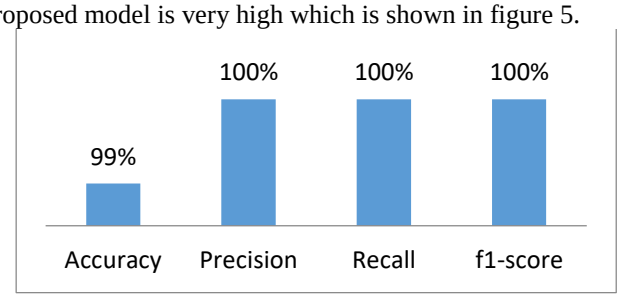

Fig. 5. Results for detecting spoof attacks

VII. CONCLUSIONS

Voice spoofing detection is a crucial task in today's world where security is of utmost importance. In recent years, many machine learning models have been proposed to detect voice spoofing; one such model is the Long Short-Term Memory (LSTM) model.

To detect voice spoofing using the LSTM model, the model is trained on a large dataset of genuine and spoofed voice recordings. The model is then able to differentiate between the two types of recordings by identifying patterns in the data.

Overall, the LSTM model has shown high accuracy in detecting voice spoofing of 99% and obtained good performance metrics of the proposed model such as the precision of genuine is 1.00 and for spoof is 0.80, recall of genuine is 1.00 and for spoof is 0.89 and f1-score of genuine is 1.00 and for spoof is 0.84.

REFERENCES

[1] M. Dua, C. Jain, and S. Kumar, "LSTM and CNN based ensemble approach for spoof detection task in automatic speaker verification systems," J. Ambient Intell. Humaniz. Comput., vol. 13, no. 4, pp. 1985–2000, 2022.

[2] J. Zhou, T. Hai, D. N. A. Jawawi, D. Wang, E. Ibeke, and C. Biamba, "Voice spoofing countermeasure for voice replay attacks using deep learning," *J. Cloud Comput. Adv. Syst. Appl.*, vol. 11, no. 1, 2022.

[3] A. G. S. Sandipan Chakroborty, "Improved Text-Independent Speaker Identification using Fused MFCC & IMFCC Feature Sets based on Gaussian Filter," *Journal of Electronics and Communication Engineering*, vol. 3, no. 11, pp. 3–11, Nov. 2009.

[4] B. Chettri, D. Stoller, V. Morfi, M. A. M. Ramírez, E. Benetos, and B. L. Sturm, "Ensemble models for spoofing detection in automatic speaker verification," in *Interspeech 2019*, 2019.

[5] M. A. Hossan, S. Memon, and M. A. Gregory, "A novel approach for MFCC feature extraction," *2010 4th International Conference on Signal Processing and Communication Systems*, 2010.

[6] H. Muckenhirn, M. Magimai-Doss, and S. Marcel, "End-to-End convolutional neural network-based voice presentation attack detection," *2017 IEEE International Joint Conference on Biometrics (IJCB)*, 2017.

[7] H. Dawood, S. Saleem, F. Hassan, and A. Javed, "A robust voice spoofing detection system using novel CLS-LBP features and LSTM," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 9, pp. 7300–7312, 2022.

[8] Z. Xie, W. Zhang, Z. Chen, and X. Xu, "A comparison of features for replay attack detection," *J. Phys. Conf. Ser.*, vol. 1229, no. 1, p. 012079, 2019.

[9] A. Pandey and D. Wang, "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[10] Y. Zhang, F. Jiang, and Z. Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Process. Lett.*, vol. 28, pp. 937–941, 2021.

[11] N. Medikonda and S. I, "Mimicry voice detection using convolutional neural networks," in *2020 International Conference on Smart Electronics and Communication (ICOSEC)*, 2020.

Medikonda Neelima received her B.Tech. degree in Electronics & communication Engineering from M.L.E.C. Engineering college, Singarayakonda, M.E. degree in Electronic Instrumentation from Andhra University, Visakhapatnam, and pursuing Ph.D at JNTU, Kakinada. At present, She is working as an Assistant Professor at GVP college of Engineering(A), Visakhapatnam. She has 15 years of teaching experience. Her areas of interests are Speech Signal Processing and Image Processing.

Dr I. Santi Prabha is Professor in ECE departmentof JNTUK. She did her B.Tech & M.Tech with specialization In Instrumentation and Control Systems from JNTU College of engineering, Kakinada. She was awarded with Ph.D. in Speech signal processing by Jawaharlal Nehru Technological University in 2005. She has more than 30 years of experience in teaching. She is a member of various professional organizations like Fellow of Institution of Engineers (India), Fellow member of The Institution of Electronics and Telecommunication Engineers and Life member of Indian Society for Technical Education. She has published more than 90 technical papers in National and International journals/ conferences. She worked as Head of ECE Department and subsequently served as Director of EOW&G, JNTUK, Kakinada during 2009 – 2015. She served as Rector, JNTUK, Kakinada during 2018-2019.