International Journal of Advanced Engineering and Management System
Volume 2, Issue 3, July – September (2026) Arrived: 23/07/2026 : Accepted: 03/08/2026

RESEARCH ARTICLE

# Deep Learning Based Feature Level and Image Level Framework for Fake voice Detection

S. Sharanyaa[1], Shalini S[2], Madhumitha P[3], Sherlin Jemima J[4]
[1,2,3,4]Department of Information Technology, Panimalar Engineering College, Chennai
[1]sharanyaasigamani@gmail.com

***Abstract*- This paper introduces a new framework for the identification of artificial (Deep Fake Audio) voice using both the feature level and image level learning approaches in deep learning based technique. The rapid development in Artificial Intelligence and generation techniques has made the creation of highly believable audio possible; this poses serious threats to digital security and credibility of the information. In order to deal with this issue, the proposed system consists of a two-branch system consisting of Mel Frequency Cepstral Coefficient (MFCC) and Gammatone Frequency Cepstral Coefficient (GFCC) with a feature level approach using a Long Short-Term Memory (LSTM) model and Mel spectrogram images with an image-level approach using Deep Convolutional Neural Network (DCNN). The system uses advanced techniques of preprocessing, feature extraction and CatBoost based feature fusion. Experiments have been performed using the ASV spoof 2021 dataset for testing the efficacy of the proposed system in discriminating between real and fake audio. The system gives better accuracy, precision and robustness as compared to some existing state-of-the-art techniques. The system can also be used for real-time detection through graphical user interface.**

***Keywords- Deepfake Audio, MFCC, GFCC, Long Short Term memory, Convolutional Neural Network, Deep Learning, Spectrogram***

## I. INTRODUCTION

In the digital age, there has been an incredible growth in Artificial Intelligence (AI), which has made it possible to create robust Generative models that can be used to create highly realistic synthetic media. Deepfake audio has proved to be a big threat due to the ability of these techniques to create convincing speeches that are similar to human voices. There are many dangers posed by these developments, including issues of privacy, security, and trustworthiness of the digital communication system. Some of the dangers posed by the misuse of deepfake audio include impersonation, financial fraud, and the propagation of disinformation and propaganda.[1].

To to tackle this problem, there is an urgent requirement of detection mechanisms that could reliably differentiate between genuine and deepfake audio. Good detection systems would help create secure applications for verifying the authenticity of audio files and reducing the effects of deepfake attacks on society and technology. Recent findings of increased number of deepfake attacks resulting in fraud cases have underscored the necessity of detection frameworks [2].

In this regard, the current study emphasizes the design of a strong deep fake audio detection system by applying machine learning and deep learning methods. This technique entails conducting an analysis on curated data sets containing both real and fake audio files so as to identify the differentiating features between these two types of speech signals. Through the use of sophisticated learning algorithms, the system will learn to detect subtle differences in deep fake audio [3].

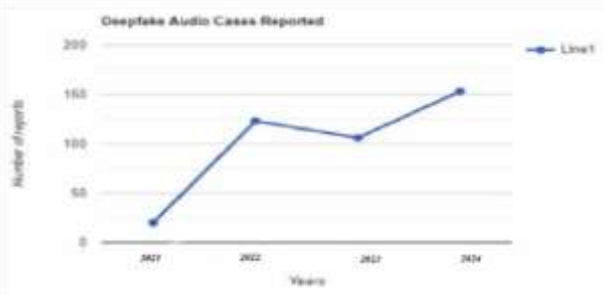


Fig 1. Deepfake Audio Cases Reported

Furthermore, This study systematically evaluates and compares different techniques for detecting deepfakes of audio to determine the best methods to classify genuine and deepfake audio signals. In addition, an implementation framework is proposed to detect audio fakes in real-time. In order to tackle the shortcomings of existing methods, the

 

proposed study considers a deep learning framework, which places emphasis on effective feature representation and model generalization [4]. Transfer learning is used to leverage the knowledge from pre-trained models for better accuracy and efficiency. Effectiveness of the proposed model is established by experiments conducted on both the custom and benchmark datasets. In addition to this, advanced feature extraction techniques, and other techniques such as Mel-frequency cepstral coefficients (MFCC) are fused with convolutional neural networks (CNN) and fully connected layers for the extraction of discriminant features of the audio signal. Hierarchical feature learning is performed by the CNN model, and the fully connected layers help achieve accurate classification. The entire framework exhibits good results in identifying deepfake audio with high accuracy and robustness.

## II. LITERATURE SURVEY

The rapid development of Artificial Intelligence (AI) and Machine Learning (ML) made possible the emergence of very sophisticated solutions for multimedia content creation and manipulation. One of those is the technology of deepfakes which proves the potential of generative models at the same time being one of the most dangerous threats regarding authenticity, privacy, and security [6]. While originally related to the creation of synthetic videos, deepfake technology has evolved into very realistic speech generation, able to replicate human voice. It brings many problems to industries like banking, politics, entertainment, and law enforcement as the use of synthetic speech can cause fraud, disinformation, and even break the security of voice biometrics [8]. While traditionally considered to be a rather safe way of authentication, voice biometrics becomes more vulnerable in connection with the progress of neural Text-toSpeech (TTS) and Voice Conversion (VC) [9]. Currently, deep learning architectures like Transformer-based models of Wav2Vec2 allow creating speech synthesis of high quality without too much input data, posing threat to security solutions [11].

To address these challenges, deepfake audio detection techniques have developed two broad categories of approaches: classical machine learning and deep learning [12]. Classical techniques depend on manually designed features such as Mel Frequency Cepstral Coefficients (MFCC) and Constant-Q Cepstral Coefficients (CQCC), together with classifier algorithms such as Random Forest and XGBoost. Although computationally efficient, these techniques are ineffective at generalizing across novel attacks [15]. On the other hand, deep learning techniques based on Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks can learn more sophisticated features from the input, hence increasing the efficiency of detection. Additionally, end-to-end models such as RawNet2 and Wav2Vec2 do not depend on hand-designed features [17]. Despite However, due to these improvements, the majority of current research work relies on dataset limitations and mainly uses English voice [18]. With the growing instances of abuse of the deepfake voice for global communication, there is an evident requirement of efficient detection techniques that can manage different accents and environments [19].

## III. PROPOSED METHODOLOGY

This research is concerned with deepfake threats to information authenticity and presents an approach to detecting fake voices based on the use of real and synthesized datasets. A dual-branch architecture combining MFCC and GFCC feature-based (LSTM) and spectrogram-based (CNN) pipelines are used in this work for robust deepfake audio classification.

The proposed framework consists of two main stages.

- In the initial phase, the audio signal is preprocessed to enhance its quality, after which key features MelFrequency Cepstral Coefficients (MFCC) and Gammatone Cepstral Coefficients (GFCC) are extracted. These feature sets are then combined through a CatBoost-based fusion approach, and the integrated representation is subsequently provided as input to a Long Short-Term Memory (LSTM) network for model training and classification.

- In the second stage, Mel spectrograms are generated from the input audio signal and provided as input to a Deep Convolutional Neural Network (DCNN) for the classification of audio samples as real or fake. The dual system architecture work flow of proposed Deepfake Audio Detection system is given in Figure 2.

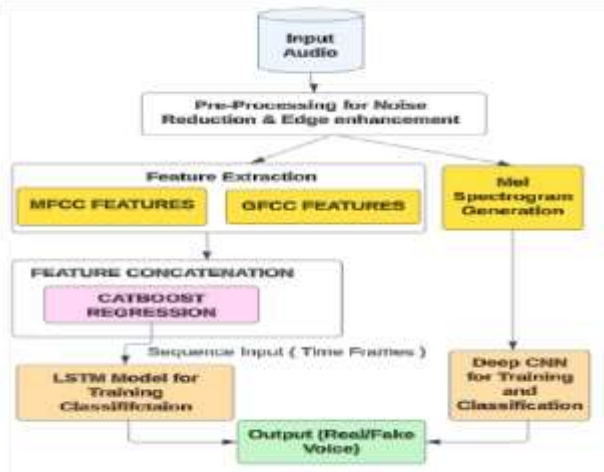


Fig. 2. Work Flow of the Proposed Deepfake Audio Detection System.

### A. Data Collection

In artificial intelligence, data quality is vital for model performance. Voice samples from a variety of sources such as public speeches and audio books have been used in this dataset

along with a balanced distribution of male and female voices as well as different age groups and accents, where a sample distribution of fake and real audio is shown in Figure 3a and 3b.

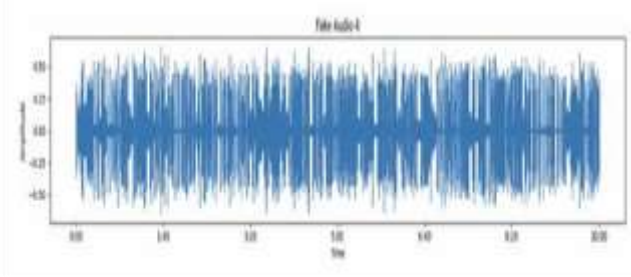

Fig.3a. Fake audio wave

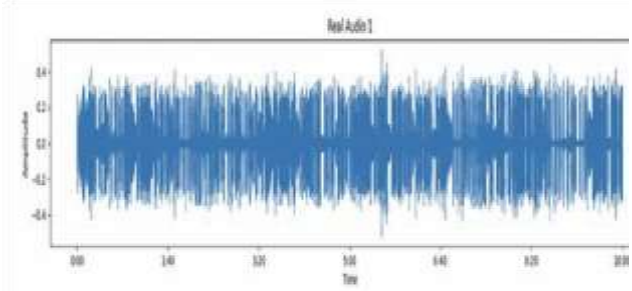

Fig.3b. Real audio

*B. Preprocessing*

The proposed method pays more attention to the preprocessing step in which the Librosa library is used to extract measurable features from the audio samples. At this stage, the speech signal is first divided into equally-sized frames to normalize the inputs. Each frame includes a specific number of samples and overlaps with the previous one by 25%, and sampling is done at a frequency of 22.05 kHz (the default value for Librosa).

*C. Feature Extraction*

This system takes raw audio samples of the prehistoric dataset of ASVspoof 2021 [20] that contain both real and spoofed speech samples. Audio normalization, silence reduction, and preprocessing take place. This entails feature extraction of the following features: MFCC and GFCCC.

***MFCC (Mel Frequency Cepstral Coefficient)***

The MFCC, Mel Frequency Cepstrum Coefficients, is a process used in audio signal processing for extracting features. The process entails different stages where the audio input signal is analyzed before MFCC extraction. In the first stage, Pre-Emphasis technique is used where the input signal is filtered to emphasize the high frequency part. This technique gives an ideal balance of the speech signal spectrum. The phonation signal $x(n)$is first divided into short-time frames and multiplied by a window function, yielding $x(n) \cdot w(n)$, where $n$ = 1,2, … , $N$ and $w(n)$represents the Hamming window. This helps to remove spectral leakage from the signal.

Following this step, the Discrete Fourier Transform is used to transform the signal from the time domain to the frequency domain and get the magnitude spectrum of the signal. The magnitude spectrum of the signal is filtered through a filter bank consisting of triangular filters in the mel-scale filter bank that approximates the human ear. Finally, the output from the filter bank is subjected to the Discrete Cosine Transform to get Mel Frequency Cepstral Coefficients.

*A. Frame Blocking and Windowing*

Windowing is an approach where a continuous sound wave is separated into equal frames, and the edges of each frame get slowly dampened to ensure smooth transitions. The common approach in use is that of the Hamming window, which eliminates any sharp changes in the edges and makes the wave smoother. This helps to minimize distortions while applying the Discrete Fourier Transform. The combination of frame blocking and windowing helps eliminate any discontinuities and noise, ensuring accurate results of spectral analysis.

*B. Discrete Fourier Tranform*

The Discrete Fourier Transform is used to map the time domain input speech signal into the frequency domain with the use of a set of spectral coefficients. The result of this mapping is a magnitude spectrum of the signal that depicts a well-defined frequency distribution of the signal components.

$$H(v) = \sum n = 0M - 1x(n) \cdot e - j2\pi nv/N, 0 \leq v \leq M - 1 \quad (1)$$

*C. Mel Spectrum*

The Mel spectrum is derived from applying a Fourier Transform to the incoming signal and then running the signal frequencies through a set of filters known as the Mel Filter Bank. Mel refers to a scale based on how human beings perceive sound frequency and correlates the actual frequency with how the human ear perceives the frequency. This is expressed in the formula below,

$$QMel = 2595 \cdot log\ 10(1 + Q100) \quad (2)$$

*D. Discrete Cosine Transform*

The resulting Mel frequency coefficients undergo a Discrete Cosine Transform (DCT) to produce a set of cepstral coefficients. It's important to note that the Mel spectrum is represented on a logarithmic scale in this process.

$$c(n) = \sum m = 0M - 1log\ 10(s(m)) \cdot cos\ (\pi n(m - 0.5)M), n = 0,1,2, \dots, C - 1 \quad (3)$$

Here, $c(n)$ denotes the cepstral coefficients, while $C$represents the total number of MFCC features extracted. In this work, a total of 39 MFCC coefficients were calculated, which also include the energy-related coefficients.

***GFCC (Gammatone Frequency Cepstral Coefficient)***

Gammatone-based feature extraction is employed to analyze the response of auditory filters in an audio signal. The gammatone filter itself is derived from its impulse response, which is modeled using a Gamma distribution function. The center frequency, representing the tonal components of the input audio, is calculated using the formula provided below,

$$G(t) = K \cdot t(n-1) \cdot e - 2\pi fct \cdot cos\ (2\pi fct + \phi)(4)$$

Here, $G(t)$ represents the impulse response of the gammatone filter, $K$ denotes the amplitude factor (peak value), and $n$ indicates the filter order. The parameter $fc$corresponds to the center frequency in Hertz, Φrepresents the phase shift, and $B$defines the duration of the impulse response. The gammatone filter is constructed as a cascade of multiple filters within a gammatone filter bank. After obtaining the filter responses, a logarithmic operation is applied, followed by the use of the Discrete Cosine Transform (DCT) to capture perceptual characteristics such as the loudness of the audio signal. The resulting coefficients, known as GFCCs, are computed as defined in Eq(5).

$$GFCC(m) = 2N\sum n = 1Nlog\ (Xn) \cdot cos\ (\pi n(m - 12)N) \quad (5)$$

In this regard, $m$ is between 1 and $M$, in which $M$ represents the number of GFCC coefficients. $Xn$ represents the energy of the $n$ th spectral band, while $N$represents the number of gammatone filters used in the filter bank. In this research, 39 GFCC coefficients have been used as the main feature set. Figure 4 shows the entire process of extracting features by both MFCC and GFCC techniques.

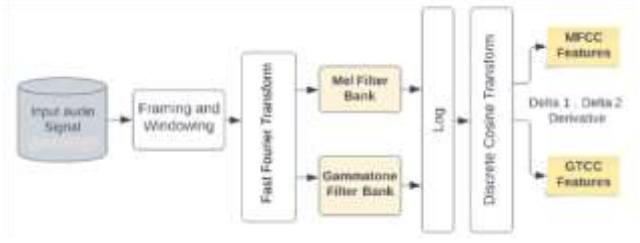


Fig. 4. Blocks involved in MFCC & GFCC Feature Extraction

*E. Feature Mapping*

In the proposed method, a feature-driven analysis In this way, a framework is implemented whereby the MFCC and GFCC features have been combined into one single unified feature vector. This combined feature vector is formed into a feature map structure for further processing. The combined vector, which comprises unique features from the input audio data, is then subjected to cat boost regression for correlation analysis.

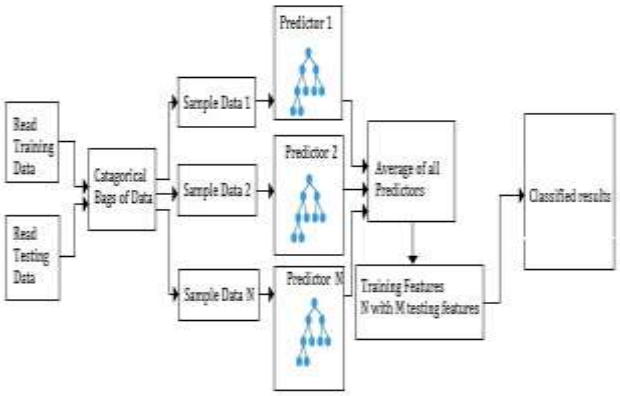


Fig. 5. CatBoost Regression for feature mapping

The feature selection process is carried out after feature extraction for the purpose of identifying the most important features of the input signal. The feature selection process is done before the cat boost regression using the adaptive process of introducing statistically derived features as feature vectors, thus avoiding the manual process of selecting them. Table 1 below indicates that the feature mapping was carried out through the use of the MFCC and GFCC coefficients with 39 each of them constituting a total of 78 features.

TABLE 1: MFCC & GFCC EXTRACTED COEFFICIENTS

| **MFCC Feature Type** | **Count** |
|---|---|
| MFCC Coefficients | 12 |
| Delta Coefficient Cepstral factor | 12 |
| Second Delta Coefficient factor | 12 |
| Energy Coefficient ,Delta & Second Delta Coefficient factor | 3 |
| Total | 39 |
| **GFCC Feature Type** | **Count** |
| GFCC Coefficients | 12 |
| Delta Coefficient Cepstral factor | 12 |
| Second Delta Coefficient factor | 12 |
| Energy Coefficient ,Delta and Second Delta Coefficient factor | 3 |
| Total | 39 |

*E. Generation of Mel Spectrogram*

Mel Spectrograms play a critical role in audio-based deepfake detection since they transform audio signals into a time–frequency domain that is well suited to analyzing speech properties. The application of Mel scale makes spectrograms consistent with the way humans perceive sounds, thus allowing for more informative analysis of sound patterns. Moreover, this type of spectrogram shows some unique spectral patterns and artifacts specific to artificial audio signals

that cannot be seen in the time domain. Finally, the Mel spectrograms can be considered image-like signals, which makes it possible to use CNNs to learn spatial patterns in audio data. Block Diagram to Generate Mel Spectrogram images is shown in Figure 6.

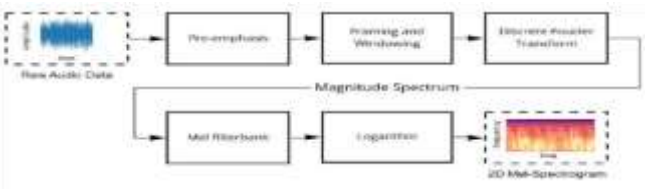


Fig. 6. Block diagram to create Mel spectrogram images from input audio signals

## IV. MODEL TRAINING AND CLASSIFICATION

### *A. Spectrogram for spatial learning using Deep Convolutional Neural Network (DCNN)*

A spectrogram is the commonly used visualization technique for audio signals, which provides visual representation of energy distribution with respect to time and frequency with the use of magnitude values of Short-Time Fourier Transform (STFT). Since the conventional spectrogram doesn't correspond to human auditory perception, the audio signal is converted into Mel Spectrogram with the help of Mel filter bank. The conversion into Mel Spectrogram allows catching the spectral artifacts and patterns that can help in differentiating real and fake audio.

The obtained Mel Spectrogram is split into training and test datasets and then presented to deep Convolutional Neural Network (CNN) for classification. The 2D convolutional layer with kernel size 3×33 times 33×3 is used in order to extract features and Rectified Linear Unit (RELU) activation is applied. In order to reduce dimensions of the data and increase robustness of features max pooling is performed. After that the obtained features are passed to fully connected layers for classification of the input image into the categories of real and fake audio. The overall architecture of DCNN is shown in Figure 7.

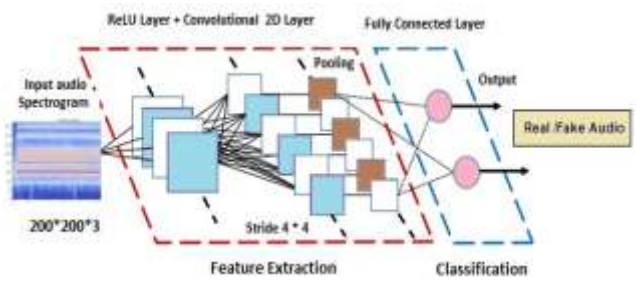


Fig.7. Deep Convolutional Neural Network (DCNN) Architecture for Fake Audio Detection

### *B. Long-short term memory for temporal learning*

For deepfake audio detection, the MFCC and GFCC feature vectors are calculated for each frame in 39 dimensions. Whereas MFCC features are obtained by taking into consideration the perceptually relevant frequencies based on the Mel scale, the GFCC features can better model the human auditory system through the use of Gammatone filters. These two sets of features are then concatenated together to obtain a feature vector of 78 dimensions per frame.

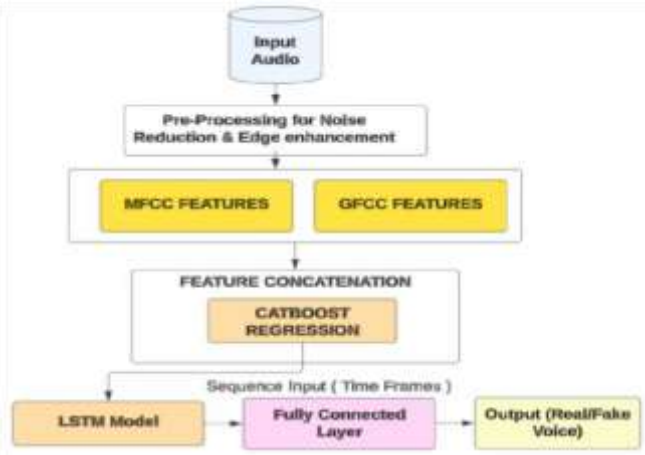


Fig. 8. Temporal Deepfake Audio Detection Using Fused MFCC and GFCC Features with LSTM

The sequence of the fused feature vectors across time frames is then utilized as the input for the Long Short-Term Memory (LSTM) model. This works well because LSTM is designed in such a way that it can learn temporal dependencies and sequences from speech signals. Through temporal variations analysis, the LSTM will be able to distinguish inconsistencies and artifacts in the generated audio that could not easily be detected from frame-level analysis. The last hidden states of LSTM are passed to the fully connected layers and sigmoid activation function is applied to produce the final result, where the input is classified as real or fake audio. The following diagram depicts the Feature Level Fusion of MFCC and GFCC with LSTM for Deepfake Audio Detection

## V. EXPERIMENTAL SETUP

Python 3.9 was employed to realize the proposed framework, using libraries like TensorFlow, Keras, PyTorch, Librosa, and Scikit-learn. The Librosa library helped in feature extraction in the form of MFCC and CQCC. Training and testing was done by using GPU enabled TensorFlow on ASVspoof 2021.

## VI. RESULTS AND DISCUSSION

The suggested deep fake audio detection technique has the capability to detect faked audio under diverse acoustic scenarios and background noise. The use of MFCC feature extraction ensures that spectral characteristics of speech are uniformly represented. A CNN-LSTM architecture improves detection performance through the combination of convolutional spatial feature generation and LSTM temporal sequence prediction. The proposed dual-branch network with a combination of feature-based (MFCC-GFCC + LSTM) and

spectrogram-based (CNN) networks for effective classification of deepfake audio. Probabilistic classification ensures accurate classification of real vs fake audio with high accuracy. It performs better than individual CNN, RNN, and LSTM networks while maintaining computational efficiency.

### *A. Graphical User Interface Output Recognition of Real and Fake Voice using Mel-Spectrogram*

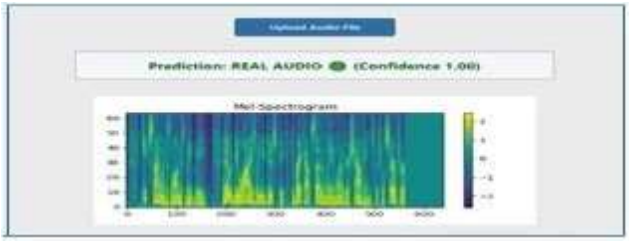


Fig.9. GUI output showing successful classification of real audio with confidence score 1.00

In the case of a real audio sample, the Graphical User Interface (GUI) will identify the type as REAL AUDIO with a very high confidence score of 1.00, as demonstrated in Figure. 9 below. The identification of such a real audio sample is made evident by the use of a green warning box on the Graphical User Interface (GUI). The Mel-Spectrogram generated for such a sample will have a regular pattern since there will be stable speech harmonics over time. This is an indicator that the audio sample contains real prosodic variations. In a similar manner, in the case where the audio sample has been altered or synthesized, the system will classify it as FAKE AUDIO, as shown in Figure 10 below. This identification is made at a high confidence rate of 1.00, which is indicated by a red color in the GUI.

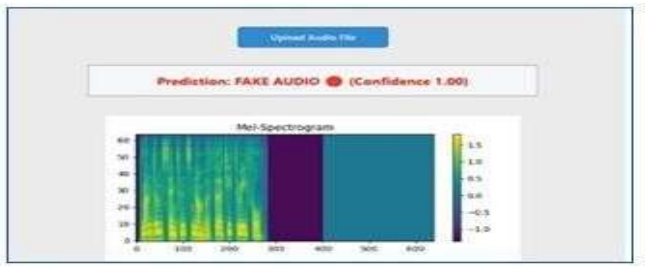


Fig. 10. GUI output showing detection of fake audio with confidence score 1.00.

### *B. Feature Extraction Results*

The simulation of audio signal processing stages for both real and synthetic audio signals is illustrated in Figure 11. The input voice signals for both the real and synthetic audio along with their respective MFCC and GFCC feature representations are provided here, showing the differences in their spectral/perceptual characteristics. These features are then fed as input to the LSTM classifier for discriminating between real and synthetic audio.

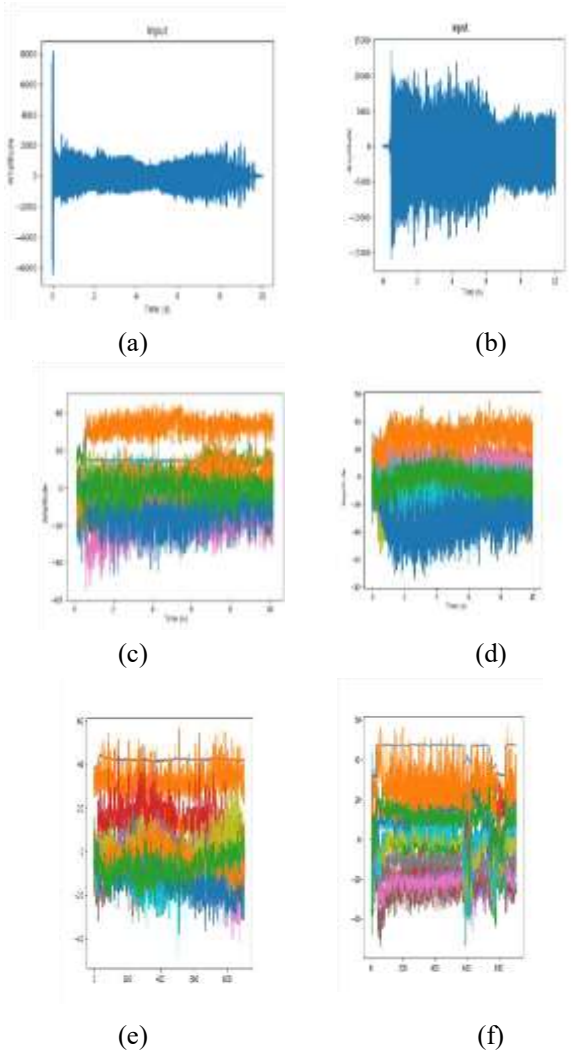


Fig. 11. Simulation of various Audio processing steps a) Input voice signal of Real Audio, b) Input voice signal of Fake Audio, (c) MFCC of Real Audio, (d) MFCC of Fake Audio (e) GFCC of Real Audio (f) GFCC of Fake Audio.

### *C. Performance Metrics*

The Performance of the suggested model is calculated through parameters like Accuracy, Sensitivity, Specificity, and Matthews Correlation Coefficient (MCC). In this regard, $Tp$is true positive, $Tn$is true negative, $Fp$is false positive, and $Fn$is false negative. Accuracy is determined based on the comparison of predicted output and the actual result, and its formula is provided in Eq(6).

$$Acc = Tp + TnTp + Tn + Fp + Fn \quad (6)$$

Sensitivity Evaluates the capacity of the suggested system in identifying PD patients through classification outcomes, as defined in Eq(7).

$$Sen = TpTp + Fn \quad (7)$$

Specificity is the measure of the proportion of genuinely healthy cases that have been correctly classified as healthy in Eq(8).

$$Spec = TnTn + Fp \quad (8)$$

The Mathews Correlation Coefficient (MCC) is used to compute the correlation between the predicted and

actual outputs, according to Equation (9). The range of MCC is from -1 to +1, where the closer the MCC to +1, the better the correlation between predicted outputs and actual labels. The MCC value of +1 shows complete correlation, whereas the value of -1 is indicative of complete disagreement between predicted outputs and actual labels.

$$MCC = Tn \cdot Tp - Fp \cdot Fn(Tn + Fp)(Tp + Fn)(Tn + Fn)(Tp + Fp) \quad (9)$$

*D. Model Performance Comparative Analysis:*

The The suggested system achieved optimal performance and also showed good generalization on new data. This is because the hybrid CNN and LSTM architecture, which uses temporal and spectral feature extraction, learns the subtle variations in audio that cannot be captured by the traditional CNN architecture. The design of the GUI makes the model user-friendly.

A The higher the correlation score, the better the performance of the CatBoost Regression algorithm. The proposed CBR model obtains the average value of MCC equal to 0.99 when trained on the proposed MFCC and GFCC features combination, indicating the efficiency of the model. They can be considered reliable statistical performance measures. In figure 12 the results of MCC after application of CatBoost regression are presented.

It should be noted that despite the lower recall of the LSTM model (92.1%) comparing to its precision, the difference between these measures can be considered insignificant, which indicates balanced performance of the model. Higher precision value indicates that the model is highly accurate and reliable for detection of fake audio with low false positives.

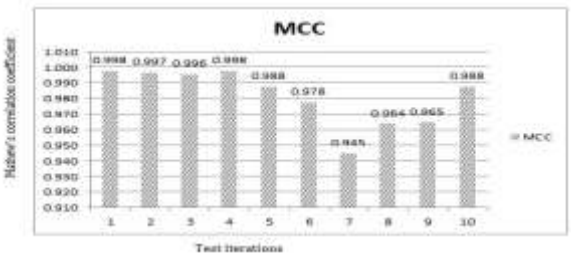


Fig. 12. Mathew's correlation coefficient after CBR mode

The Performance comparison of the DCNN model trained on the Mel-spectrogram features and the LSTM model trained on the combination of MFCC and GFCC features shows the efficiency of the proposed method. Although the DCNN model reaches an excellent result and gains an accuracy rate of 89.8%, thus showing its ability to learn spatial patterns based on spectrograms, the performance is relatively poor compared to that of the LSTM model. On the other hand, the LSTM model significantly exceeds the performance of the DCNN model and reaches an accuracy rate of 98.7% and a precision of 98.4%. This can be explained by the fact that the LSTM architecture is able to learn the time dependencies of the sequential audio features, which are important for distinguishing real and fake audio signals.

The greater precision implies that the model is extremely accurate in detecting fake audios with very few false alarms, an important aspect in deepfake detection systems. On the whole, the results indicate that combining perceptually-relevant features and the use of sequence modeling produces better results than traditional spectrogram-based CNN systems.

TABLE 2: PERFORMANCE COMPARISON OF PROPOSED FEATURE LEVEL AND IMAGE LEVEL LEARNING FRAMEWORK FOR FAKE VOICE DETECTION

Table 2 compares the performance of the two frameworks for detecting fake voices based on Mel-spectrogram, MFCC, and GFCC features. Table 3 compares the performance of the system proposed against different state-of-the-art systems evaluated on different datasets.

| Model | Accuracy( %) | Precision( %) | Recall (%) |
|---|---|---|---|
| DCNN using Mel-Spectrogram (Image Level) | 89.8 | 92.2 | 91.9 |
| LSTM using MFCC and GFCC Features (Feature Level) | 98.7 | 98.4 | 92.1 |

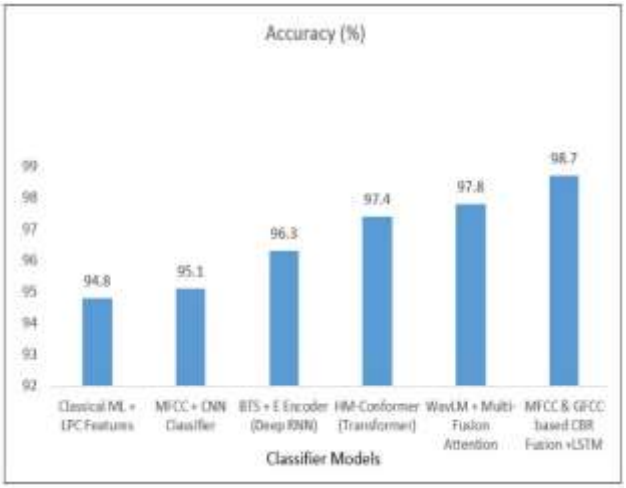


Fig.13. Comparison of the Proposed Model with state of art techniques in terms of Performance

## VII. CONCLUSION

This works present an efficient deepfake audio detection method that uses feature level and image level learning mechanisms. Inclusion of MFCC and GFCC features along with LSTM makes temporal analysis more efficient, and Mel spectrograms along with CNN helps in capturing the spatial components of audio data. This two-branch model has proven to be superior in terms of detection accuracy and precision than any single model. Experiments show that the proposed model has high accuracy and precision rates, along with good generalization capabilities. Moreover, CatBoost-based feature fusion makes the system more discriminative.

Moreover, the graphical user interface developed showcases the practicality of the model in real-world scenarios. Further research efforts may involve making the model scalable and adaptable to multilingual and noisy environments. Inclusion of transformer architectures and self-supervised learning approaches may help improve performance of the system. Multimodal deepfake detection that combines audio and visual modals may become the next step of the research.

## REFERENCES


[1] [Multimodaltrace: Deepfake Detection using Audiovisual Representation Learning,Muhammad Anas Raza Khalid Mahmood Malik,Oakland University,2023 IEEE/CVF Conference on Computer Vision and Pat-tern Recognition Workshops (CVPRW) — 979-83503-0249-3/23/31.00©2023IEEE—DOI: 10.1109/CVPRW59228.2023.00106

[2] C. Stupp, “Fraudsters used Ai to mimic CEO’s voice in unusual cybercrime case,” Wall Street J., vol. 30, no. 8, pp. 12, 2019.

[3] Deepfake audio detection by speaker verification,Alessandro Pianese,Davide Cozzolino, Giovanni Poggi and Luisa Verdoliva ,University Federico II of Naples, Italy,2022 IEEE International Workshop on Information Forensics and Security (WIFS) — 979-83503-0967-6/22/31.00©2022IEEE—DOI: 10.1109/WIFS55849.2022.9975428.

[4] Deepfake detection using deep learning methods: A systematic and comprehensive review , Arash Heidari, Nima Jafari Navimipour, Hasan Dag, Mehmet Unal, 2023,https://doi.org/10.1002/widm.1520.

[5] D. Garg and R. Gill, ”Deepfake Generation and Detection - An Exploratory Study,” 2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), Gautam Buddha Nagar, India, 2023, pp. 888 893, doi: 10.1109/UPCON59197.2023.10434896.

[6] R. Mahyavanshi, C. V. M. Reddy, A. J. Shah and H. A. Patil, ”Teager Energy Cepstral Coefficients for Audio Deepfake Detection,” 2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Macau, Macao, 2024, pp. 1 6, doi: 10.1109/APSIPAASC63619.2025.10848893.

[7] I. Altalahin, S. AlZu’bi, A. Alqudah and A. Mughaid, ”Unmasking the Truth: A Deep Learning Approach to Detecting Deepfake Audio Through MFCC Features,” 2023 International Conference on Information Technology (ICIT), Amman, Jordan, 2023, pp. 511–518, doi: 10.1109/ICIT58056.2023.10226172.

[8] Y. Xie et al., ”Does Current Deepfake Audio Detection Model Effectively Detect ALM-Based Deepfake Audio?,” 2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP), Beijing, China, 2024, pp. 481–485, doi: 10.1109/ISCSLP63861.2024.10800375.

[9] Z. Hao et al., ”Integrating Spectro-Temporal Cross Aggregation and Multi-Scale Dynamic Learning for Audio Deepfake Detection,” ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 2025,pp. 1–5, doi: 10.1109/ICASSP49660.2025.10889337.

[10] Y. Guo, H. Huang, X. Chen, H. Zhao and Y. Wang, ”Audio Deepfake Detection With Self-Supervised Wavlm And Multi-Fusion Attentive Classifier,” ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Republic of, 2024, pp. CASSP48485.2024.10447923. 12702–12706

[11] Z. Jin, L. Lang and B. Leng, ”Wave-Spectrogram Cross-Modal Aggregation for Audio Deepfake Detection,” ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 2025, pp. 1–5, doi: 10.1109/ICASSP49660.2025.10890563.

[12] J. Choi, T. Kim and J. Choi, ”ID-Flow: Leveraging Voice Identity for Generalizing Audio Deepfake Detection,” 2025 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 2025, pp. 1–5, doi: 10.1109/ICCE63647.2025.10929900.

[13] K. M. Lapates, B. D. Gerardo and R. P. Medina, ”Performance Evaluation of Enhanced DCGANs for Detecting Deepfake Audio Across Selected FoR Datasets,” 2024 15th International Confer- ence on Information and Communication Technology Convergence (ICTC), Jeju Island, Korea, Republic of, 2024, pp. 54–59, doi: 10.1109/ICTC62082.2024.10827547.

[14] S.-P. Doan, L. Nguyen-Vu, S. Jung and K. Hong, ”BTS-E: Au- dio Deepfake Detection Using Breathing-Talking-Silence Encoder,” ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1–5, doi: 10.1109/ICASSP49357.2023.10095927.

[15] E. Jain and A. Singh, ”Deepfake Voice Detection Using Convolutional Neural Networks: A Comprehensive Approach to Identifying Synthetic Audio,” 2024 International Conference on Communication, Control, and Intelligent Systems (CCIS), Mathura, India, 2024, pp. 1 5, doi: 10.1109/CCIS63231.2024.10931997.

[16] H.S. Shin, J. Heo, J.-H. Kim, C.-Y. Lim, W. Kim and H.-J. Yu, ”HM CONFORMER: A Conformer-Based Audio Deepfake Detection System with Hierarchical Pooling and Multi-Level Clas- sification Token Aggregation Methods,” ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Republic of, 2024, pp. 10581–10585, doi: 10.1109/ICASSP48485.2024.10448453.

[17] L. Astrid, E. Ghorbel and D. Aouada, ”Audio-Visual Deepfake Detection With Local Temporal Inconsistencies,” ICASSP 2025- 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 2025, pp. 1–5, doi: 10.1109/ICASSP49660.2025.10889087.

[18] P. Chiddarwar, ”Real-Time Detection of AI-Generated Deepfake Audio: A Novel Approach,” 2024 IEEE 4th International Conference on ICT in Business Industry & Government (ICTBIG), Indore, India, 2024, pp. 1–5, doi: 10.1109/ICTBIG64922.2024.10911062.

[19] D. Song, N. Lee, J. Kim and E. Choi, ”Anomaly Detection of Deepfake Audio Based on Real Audio Using Generative Adversarial Network Model,” IEEE Access, vol. 12, pp. 184311–184326, 2024, doi: 10.1109/ACCESS.2024.3506973.

[20] Q. A. Samiha, M. T. Alom, R. Hossain, R. Islam and A. Sul- tana, ”Leveraging Deep Learning Methods for Detecting Deepfake Speeches,” 2025 IEEE 4th International Conference on Computing and Machine Intelligence (ICMI), MI, USA, 2025, pp. 1–5, doi: 10.1109/ICMI65310.2025.11141035.

[21] M. Pham, P. Lam, T. Nguyen, H. Nguyen and A. Schindler, ”Deepfake Audio Detection Using Spectrogram-based Feature and Ensemble of Deep Learning Models,” 2024 IEEE 5th International Symposium on the Internet of Sounds (IS2), Erlangen, Germany, 2024, pp. 1–5, doi: 10.1109/IS262782.2024.10704095.

[22] A. E. Djire´, A. Sabane´, A.-K. Kabore, R. Kafando and T. F. Bissyande´, ”Evaluating Acoustic Parameters for DeepFake Audio Identification,” 2023 IEEE Afro-Mediterranean Conference on Artificial Intelligence (AMCAI), Hammamet, Tunisia, 2023, pp. 1–6, doi: 10.1109/AMCAI59331.2023.10431521.