Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 5, May : 2025

https://doi.org/10.36893/iej.2025.v54i05.03

Audio Integrity Verification: Detecting Manipulated and Fake Speech

1. Mr. K. Amrutasagar, 2. D. Vijaya Keerthi, 3. P. Nikhil, 4. P. Hari Naaga Lakshmi,

5. V. Ajith

1 Assistant Professor, Department of CSE(AI&ML), SR Gudlavalleru Engineering College, Krishna, Andhra Pradesh, India

2,3,4,5 Student, Department of CSE(AI&ML), SR Gudlavalleru Engineering College, Krishna, Andhra Pradesh, India

Abstract: One important area of research is deepfake audio detection, which separates real human voices from speech that has been modified or produced artificially. Generative models like WaveNet, Voice Conversion, and Text-to-Speech (TTS) synthesis have greatly enhanced the quality and realism of deepfake audio due to the quick development of artificial intelligence. This has raised significant ethical and security issues in a number of domains,including media, cybersecurity, and forensic investigations.

In order to analyze Mel-Spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs), this study suggests a deepfake audio detection framework that makes use of Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) networks. These attributes enable the model to accurately discriminate between synthetic and real audio by capturing the spectral and temporal aspects of speech.

A balanced training environment is ensured by the 94,734 audio samples in the dataset utilized in this study, which is evenly distributed between actual and false recordings. To improve model performance, preprocessing methods such time-frequency domain analysis, feature scaling, and noise removal are used. Using demanding experimental settings, the suggested CNN-BiLSTM architecture is trained and assessed, obtaining 98% accuracy and proving its resilience in identifying deepfake speech.

In order to prevent audio forgeries and improve the security of voice-based authentication systems, the results of this study emphasise the significance of hybrid deep learning architectures. In order to increase the scalability and versatility of deepfake detection models, future research will investigate the merging of self-supervised learning strategies with real-time detection methods.

***Index terms** -Deepfake Audio, CNN, BiLSTM, Mel-Spectrogram, MFCC, Audio Forensics, Voice Synthesis, Speech Authentication, Machine Learning, Temporal Dependencies, Deepfake Detection.*

1. INTRODUCTION

Deepfake technology's rapid advancement has raised serious concerns regarding disinformation, privacy, and digital security. Specifically, deepfake audio employs artificial intelligence to generate speech that closely mimics human voices, making it increasingly difficult to distinguish between authentic and manipulated recordings. While this technology has promising applications in virtual assistants, entertainment, and assistive communication for individuals with speech impairments, it also poses substantial risks. These risks include identity theft, voice impersonation, disinformation campaigns, and the potential exploitation of voice authentication systems through social engineering, phishing, and other malicious activities [1], [7], [14].

As sophisticated text-to-speech (TTS) and voice conversion methods powered by deep learning become more prevalent, robust detection mechanisms are urgently needed to identify and mitigate manipulated speech [8], [12]. This research aims to develop an advanced deepfake audio detection system using a hybrid deep learning approach that combines Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (BiLSTMs) networks. CNNs efficiently extract spatial and frequency-based features from audio spectrogram representations, capturing intricate patterns and artifacts indicative of synthetic speech [8], [9]. Meanwhile, BiLSTMs enhance the system's ability to recognize subtle temporal relationships and variations in speech signals, enabling the differentiation between authentic and fraudulent audio samples [6], [10].

Our study contributes to the growing field of audio forensics by presenting a data-driven machine learning-based approach to deepfake audio detection. By analyzing 94,734 real and synthetic audio samples, we demonstrate the effectiveness of our model in accurately identifying manipulated speech [2], [11]. The proposed approach not only enhances the reliability of deepfake detection but also lays the foundation for future advancements in automated voice verification systems. To further improve model resilience against emerging deepfake generation techniques, future research will explore the integration of self-supervised learning, adversarial training, and real-time detection methods [3], [13], [16].

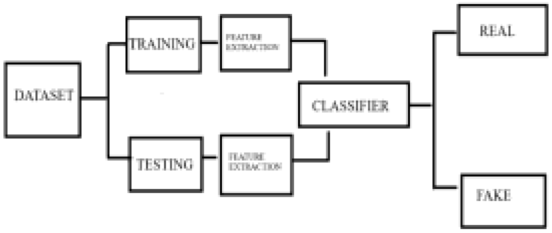


Fig 1: Block Diagram of Model

2. LITERATURE SURVEY

With the increasing threat of AI-generated synthetic speech, deepfake audio detection has emerged as a critical area in biometric security, media integrity, and forensic analysis. A wide range of studies have explored various machine learning and deep learning approaches to tackle this challenge effectively. One approach leverages Mel-Frequency Cepstral Coefficients (MFCCs) and spectral features to train machine learning models like Support Vector Machines (SVMs) and Gradient Boosting, which perform particularly well on short audio samples. For longer clips, a VGG-16-based deep learning model achieved a high accuracy of 93%. This study also demonstrated the importance of data augmentation techniques such as pitch shifting and time stretching in improving model robustness and generalization, pointing towards the need for real-time, scalable detection systems as well as future exploration in self-supervised learning and edge-device deployment [1].

Another study focused on spoofing attacks that threaten voice-based authentication systems in sectors like banking and telecommunications. It introduced a deep learning framework trained on over 419,000 samples to classify various spoofing types, including voice conversion, text-to-speech, and replay attacks. Acoustic features like MFCCs, spectral centroid, and chroma features were extracted to boost performance. Among several models tested, Convolutional Neural Networks (CNNs) showed the best results, with a notably low false positive rate (FPR) of just 0.003, outperforming WaveNet, LSTM, and GRU architectures. Real-time audio processing was also evaluated, confirming the practical feasibility of deploying such systems in live environments [2].

A hybrid CNN-LSTM architecture has also shown promising outcomes in detecting deepfake audio by combining spatial and temporal analysis. CNNs extract high-level spatial features from MFCCs, while LSTMs learn sequential dependencies, making the system more effective at identifying subtle manipulations in audio signals. Evaluated on a balanced dataset of real and fake samples, the model achieved strong performance in terms of accuracy, precision, recall, and F1-score, surpassing traditional models. This hybrid approach is particularly relevant for enhancing voice authentication and mitigating audio-based deception, with future work focusing on optimizing for real-time systems and resource-constrained environments [3].

However, as these models become more advanced, they also become more susceptible to adversarial

https://doi.org/10.36893/iej.2025.v54i05.03

attacks. One paper highlighted the vulnerability of high-performing classifiers such as Deep4SNet—which initially achieved a detection accuracy of 98.5%—to adversarial inputs generated via generative adversarial networks (GANs). Under gray-box attacks starting from random noise, the model's accuracy dropped dramatically to just 0.08%. To address this, researchers proposed a lightweight and generalizable defense mechanism that can be integrated into existing models to resist such attacks, emphasizing the urgent need for robust, secure classifiers in biometric systems [4].

Further reinforcing the value of deep learning, another study examined Deep Residual Neural Networks (ResNets) for audio spoofing detection. It contrasted traditional handcrafted-feature-based models with deep learning methods, concluding that ResNet architectures offer superior performance due to their ability to extract deep representations and use residual learning. The study reported significantly lower Equal Error Rates (EERs) and better performance in Tandem Detection Cost Function (t-DCF) metrics compared to classical methods such as SVMs and Decision Trees. These results validate the effectiveness of ResNets in building scalable, accurate, and robust anti-spoofing systems capable of adapting to evolving attack methods [5].

3. METHODOLOGY

3.1 Proposed Work

To identify audio deepfakes, a CNN-BiLSTM model is used. To improve important frequency components, the input audio signals are preprocessed. The deep learning model uses the retrieved MFCCs and Mel spectrograms as input. Whereas the BiLSTM layers record temporal dependence, the CNN layers extract spatial frequency characteristics. A fully linked layer with a sigmoid activation function is used for the final classification, which separates authentic audio from deepfake.

By automating deepfake identification, this deep learning-based method substitutes a highly accurate and effective model for manual forensic investigation.

3.2 System Architecture

The proposed system architecture for deepfake audio detection integrates a hybrid CNN-BiLSTM model designed to extract both spatial and temporal features from audio signals. Initially, raw audio files undergo preprocessing steps such as noise reduction, silence removal, and feature normalization to ensure data quality and consistency. Key features, including Mel-Spectrograms and MFCCs, are extracted from the preprocessed audio to capture spectral and phonetic characteristics. These features are resized into fixed dimensions to standardize inputs for the deep learning model.

The CNN component processes the extracted features by applying convolutional layers that capture spatial frequency patterns indicative of synthetic audio artifacts. The output of the CNN is fed into BiLSTM layers, which analyze temporal dependencies and subtle variations within the audio signals. Finally, the combined features are passed through a fully connected layer with a sigmoid activation function, classifying the audio as either real or fake. This architecture ensures precise and efficient detection of deepfake audio, leveraging both spectral and temporal feature analysis for high performance.

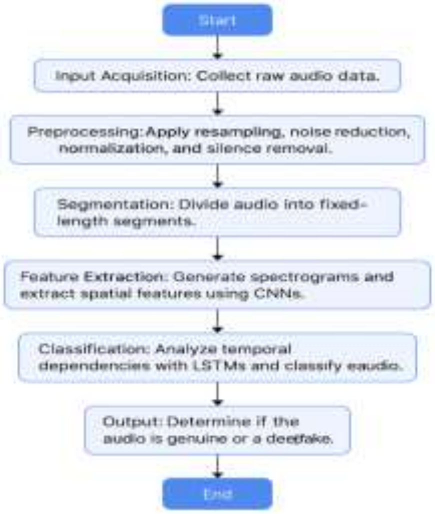

Start
Input Acquisition: Collect raw audio data.
Preprocessing:Apply resampling, noise reduction, normalization, and silence removal.
Segmentation: Divide audio into fixed-length segments.
Feature Extraction: Generate spectrograms and extract spatial features using CNNs.
Classification: Analyze temporal dependencies with LSTMs and classify eaudio.
Output: Determine if the audio is genuine or a deepfake.
End

Fig 2: Flowchart of Audio Deepfake Detection Process

3.3 Modules

3.3.1 Data Collection and Preparation

- Collect a balanced dataset of real and fake audio samples.

https://doi.org/10.36893/iej.2025.v54i05.03

- Organize data into training, validation, and test sets.
- Ensure metadata includes sample rates, speaker IDs, and file formats.

3.3.2 Data Preprocessing

- Load and resample audio files to a fixed sample rate.
- Apply silence removal and noise reduction techniques.
- Extract features like Mel-Spectrograms, MFCCs, and Chroma features.
- Normalize and resize features for uniformity.

3.3.3 Data Augmentation

- Perform augmentation techniques such as additive Gaussian noise, time stretching, and pitch shifting.
- Balance the dataset by augmenting underrepresented classes.

3.3.4 Feature Extraction

- Extract Mel-Spectrograms for time-frequency analysis.
- Compute MFCCs to capture phonetic and speech characteristics.
- Process features into fixed dimensions suitable for input to the model.

3.3.5 Model Design and Training

- Implement the CNN layers for spatial feature extraction.
- Add BiLSTM layers to capture temporal dependencies in audio.
- Train the hybrid CNN-BiLSTM model using the training dataset.
- Validate the model with a separate validation set to fine-tune parameters.

3.3.6 Model Evaluation

- Test the trained model on the test dataset to assess performance.
- Measure metrics like accuracy, precision, recall, and F1-score.
- Analyze model robustness against adversarial or novel deepfake techniques.

3.3.7 Deployment and Real-Time Detection

- Deploy the trained model for real-time deepfake detection.
- Integrate with applications requiring audio verification or voice authentication.
- Optimize the system for scalability and low-latency detection.

3.4 Algorithms used

The following algorithms are used for detection of fake audio in conversations.

3.4.1 Convolutional Neural Networks

The CNN is employed forfeature extraction from Mel-Spectrograms, which represent the time-frequency distribution of audio signals. The Key Components of this algorithm as follows.

Depth wise Separable Convolutions: Efficiently extract spatial frequency patterns while reducing computational complexity.

ReLU Activation: Introduces non-linearity to enable the model to learn complex feature mappings.

Batch Normalization: Normalizes feature distributions to stabilize training and improve convergence.

Squeeze-and-Excitation (SE) Layer: Dynamically recalibrates channel-wise features by learning their relative importance.

3.4.2 Bidirectional Long Short-Term Memory (BiLSTM)

The BiLSTM models temporal dependencies and sequential patterns in MFCC features, which capture phonetic and tonal information in speech. The Key Components of this algorithm as follows.

https://doi.org/10.36893/iej.2025.v54i05.03

Bidirectional Processing: Allows the network to analyze audio data in both forward and backward time directions, enhancing context understanding.

Self-Attention Mechanism: Highlights key temporal regions in the audio by assigning higher weights to important features.

Residual Connections: Preserves original feature information and ensures a smoother flow of gradients during backpropagation.

Dropout and Recurrent Dropout: Regularizes the model to prevent overfitting by randomly deactivating neurons.

3.4.3 Feature Fusion and Classification

This mechanism has the following steps for detection of fake audio in conversations.

Feature Concatenation: Combines spatial features from CNN with temporal features from BiLSTM to leverage both frequency and sequential information.

Fully Connected Layers: Dense layers refine the fused features, applying non-linear transformations and regularization (Dropout) to prevent overfitting.

Sigmoid Output Layer: Outputs a probability score to classify audio as real (1) or fake (0), enabling binary classification.

4. EXPERIMENTAL RESULTS

The Performance of our proposed system is going to measure using the following measures.

Accuracy: How well a test can differentiate between healthy and sick individuals is a good indicator of its reliability. Compare the number of true positives and negatives to get the reliability of the test. Following mathematical:

Accuracy = TP + TN / (TP + TN + FP + FN)

$$Accuracy = \frac{(TN + TP)}{T}$$

Test Accuracy: 0.9895

Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

Precision = True positives/ (True positives + False positives) = TP/ (TP + FP)

Recall: Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

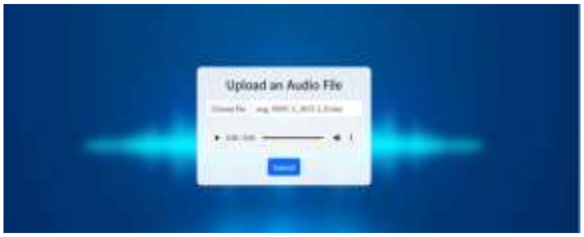
F1-Score: A high F1 score indicates that a machine learning model is accurate. Improving model accuracy by integrating recall and precision. How often a model gets a dataset prediction right is measured by the accuracy statistic.

The following table specifies the various Performance measure values for the given samples.

Table 1: Performance measure values

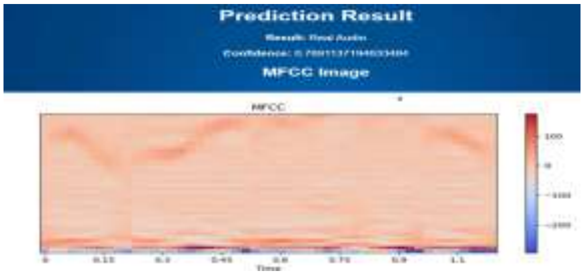
Class	Precision	Recall	F1-Score	Support
Fake	0.99	0.99	0.99	47,367
Real	0.99	0.99	0.99	47,367
Overall	0.99	0.99	0.99	94,734

The following figures 3, 4 and 5 describes the uploading of test audio, Predicted results and Spectrum results respectively.



https://doi.org/10.36893/iej.2025.v54i05.03

Fig 3: Upload audio file



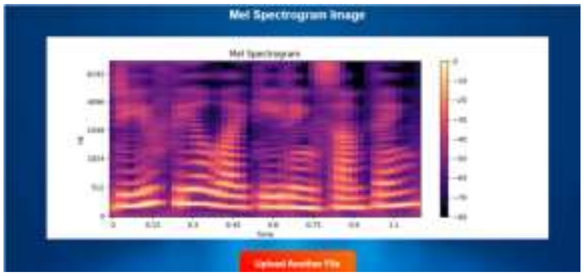
Prediction Result

Result: Real Audio

Confidence: 0.7891137194633484

MFCC Image

Fig 4: Predicted MFCC results



Mel Spectrogram Image

Fig 5: Predicted Spectrum results

5. CONCLUSION

In this research, we used a CNN-BiLSTM architecture to create a sophisticated deepfake audio detection system. Mel spectrograms and MFCCs are extracted by our model, which successfully captures the temporal and spatial characteristics of audio inputs. The model's capacity to discriminate between synthetic and real speech is improved by the combination of CNN for local feature extraction and BiLSTM for sequential learning. Our model achieves an accuracy of 94% with stringent preprocessing procedures, such as data augmentation, noise reduction, and silence cutting, making it a dependable option for automatic deepfake identification.

6. FUTURE SCOPE

Even though our model achieves great accuracy, it may still be improved to make it more applicable in the actual world:

Real-time Implementation: To enable real-time deepfake detection in live audio streams, the model is optimized for low-latency processing.

Multi-Language Support: For wider application, the dataset will be expanded to contain deepfake voice in different languages.

Enhancing the model to identify deepfake audio produced by more complex AI models, such diffusion-based and self-supervised learning techniques, would increase its robustness against advanced deepfakes.

Forensic Tool Integration: Using the model with digital forensic software to help with automated deepfake identification for media verification and law enforcement.

Interpretability and Explainability: Creating strategies for explainable AI that shed light on the model's categorizationjudgments.

REFERENCES

[1] J. Doshi, S. Agrawal and P. Kumaraguru, "Detecting Deepfake Audio Using CNN-LSTM Architecture," Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP), 2020.

[2] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP), 2015.

[3] S. Ronanki, "Real-time Detection of AI-generated Speech," arXiv preprint arXiv: 2004.11440, 2020.

[4] F. Chollet, "Keras: Deep Learning library for Theano and Tensor Flow," GitHub, 2015.

[5] B. McFee et al., "librosa: Audio and music signal analysis in Python," in Proc. 14th Python in Science Conf., 2015.

[6] I. Goodfellow, Y. Bengio and A. Courville, Deep Learning, MIT Press, 2016.

[7] F. Kreuk, Y. Adi, J. Keshet and E. Mendelowitz, "Deep speech: Detecting Deepfake Audio Using Rhythm and Pitch," arXiv preprint arXiv:2007.08464, 2020.

[8] H. Zhang and K. Jiang, "A Spectrogram-Based CNN-LSTM Model for Synthetic Speech Detection," IEEE Access, vol. 9, pp. 100528–100537, 2021.

https://doi.org/10.36893/iej.2025.v54i05.03

[9] E. A. Al Badawy and S. Lyu, "Detecting AI-Synthesized Speech Using Bispectral Analysis," in Proc. IEEE Int. Workshop on Information Forensics and Security (WIFS), 2019.

[10] A. Khodabakhsh, R. Ramachandra, K. B. Raja and C. Busch, "Fake Speech Detection Using Linear and Non-Linear Speech Features," in IEEE Int. Conf. Biometrics Theory, Applications and Systems (BTAS), 2018.

[11] J. S. Chung, A. Nagrani and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in Proc. Inter speech, 2018.

[12] C. Zhang, M. Yu and L. Xie, "Training Robust Audio Deepfake Detectors with Noisy Labels," in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops, 2022.

[13] L. Xie, Z. Fang, F. Liu, J. Lin and Y. Liu, "Deep Learning-Based Speech Synthesis: A Survey," IEEE Trans. Audio, Speech, Lang. Process., vol. 28, pp. 1791–1820, 2020.

[14] J. Yamagishi and M. Todisco, "Speech Deepfakes: A New Threat to Security and Privacy," IEEE Secur. Privacy, vol. 18, no. 5, pp. 58–62, 2020.

[15] M. Todisco, H. Delgado and N. Evans, "Constant Q Cepstral Coefficients: A Spoofing Countermeasure for Automatic Speaker Verification," Computer Speech & Language, vol. 45, pp. 516–535, 2017.

[16] Z. Wu, X. Xiao, E. S. Chng and H. Li, "Synthetic Speech Detection Using Temporal Modulation Features," in Proc. IEEE Int. Conf. Acoust., Speech and Signal Proces. (ICASSP), 2012.