International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2025
Page No: 4320- 4324
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0632

AI Based – Deepfake Voice Detection System

M Barsha[1], S Charukesi[2], G K Jeevadharsini[3], R Parthiban[4]
[1,2,3]UG Scholar, Dept. of CSE, Erode Sengunthar Engineering College, Tamil Nadu, India.
[4]Associate Professor, Dept. of CSE, Erode Sengunthar Engineering College,Tamil Nadu, India.
Emails: *barshamohan.197gmail.com[1], charusana1804@gmail.com[2], jeevadharsini1412@gmail.com[3], parthiban18121998@gmail.com[4]*

Abstract
The rise of artificial intelligence has enabled the creation of deepfake technologies that can generate highly realistic synthetic voices. While these technologies have positive applications in entertainment and accessibility, they also pose serious threats such as impersonation, misinformation, and voice-based fraud. Detecting fake or manipulated audio has therefore become an essential area of research in artificial intelligence and cybersecurity. This project presents a general fake audio detection system capable of identifying not only AI-generated deepfake voices but also manipulated real recordings such as splicing, pitch modification, and voice cloning. Unlike existing models that are limited to specific languages or static audio files, the proposed system supports multilingual audio, including major Indian languages, and is designed to work in real-time, even during live calls.

Keywords: *Deepfake Audio, Voice Spoofing, AI Voice Detection.*

1. Introduction

Deepfake technology has rapidly advanced into a powerful tool capable of generating synthetic voices that closely resemble real individuals, accurately mimicking tone, accent, pitch, and speaking style with extraordinary precision. While this innovation supports positive applications in entertainment, dubbing, and accessibility, it has simultaneously emerged as a major cybersecurity threat. Deepfake audio is increasingly misused for fraudulent phone calls impersonating family members or bank officials, political manipulation and misinformation, and attacks on digital identity and authentication systems. These risks highlight the urgent need for robust deepfake voice detection systems capable of analyzing speech signals and identifying distortions, unnatural patterns, and inconsistencies that expose artificially generated audio. To address this challenge, the project focuses on developing an AI-driven detection system that performs speech feature extraction, converts audio into Mel-spectrogram representations, applies CNN and LSTM-based deep learning models, and conducts real-time classification to determine whether a voice is real or synthetic. The proposed model enhances the reliability of audio-based authentication and contributes to safer, more secure digital communication environments. To counter these threats, the proposed system extracts essential speech features, converts audio signals into Mel-spectrogram representations, and processes them through CNN/LSTM architectures to enable accurate classification and real-time detection. The goal of this work is to build a reliable, user-friendly deepfake audio detection mechanism that enhances digital security, improves trust in audio communication, and strengthens audio forensics in modern cybersecurity environments [1-3].

1.1 Methods of Audio Detection

Audio Preprocessing: This involves removing background noise, trimming silent parts, normalizing volume levels, and converting all audio files to the same sampling rate. These steps help ensure that the model receives clean and consistent audio signals, making it easier to compare real and fake samples.

Convolutional neural Network Layers:- CNN layers are used to analyze the visual patterns present in Mel-spectrograms and other audio feature images. They detect shapes, edges, and textures within these representations, allowing the model to identify unusual frequency patterns, distortions, or artifacts commonly found in deepfake audio. CNNs help capture the “spatial” characteristics of speech signals.

Classifier: The Softmax layer takes the model’s

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2025
Page No: 4320- 4324
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0632

learned patterns and assigns probabilities to two classes: “Real Voice” and “Fake Voice.

Training Techniques: Training techniques help improve model accuracy and prevent errors. Methods such as the Adam optimizer, cross-entropy loss function, and data augmentation allow the model to learn faster and more effectively.

Evaluation Metrics: Evaluation metrics measure how well the deepfake detection model performs. Metrics like accuracy, precision, recall, and F1-score examine the correctness of predictions, while confusion matrix and ROC curves provide a detailed analysis of true and false classifications. These metrics help determine whether the model is ready for real-world use. Tables and Figures are presented center, as shown below and cited in the manuscript.

Table 1 Experimental Input Parameters for EDM

Component	Method	Category
Audio Input	Real & Deepfake audio datasets	Input Data
Preprocessing	Noice Reduction	Data cleaning
Feature Extraction	MFCC ,Mel -Spectrogram	Signal Features
Deep Learning Model	CNN -LSTM	Model Architecture
CNN Layers	Convolution + Pooling	Spectral Pattern learning
Classifier	Softmax	Output Prediction
Evaluation Metrics	Accuracy	Performance Evaluation
Training Techniques	Adam Optimizer	Model Training
Output	Real Voice	Final Result

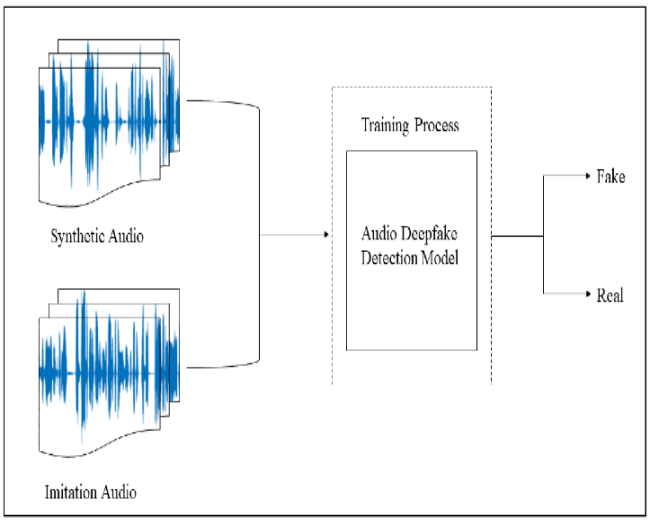

Figure 1 Confusion Matrix of Deepfake Voice Detection Model

The diagram illustrates the workflow of an Audio. The project uses real and deepfake audio samples which are first cleaned through preprocessing methods like noise removal and normalization. Important speech features such as MFCC and Mel-spectrograms are then extracted to highlight frequency and pitch patterns. A hybrid CNN–LSTM deep learning model is used, where CNN layers learn spectral features and LSTM layers learn time-based speech patterns. A Softmax classifier then predicts whether the audio is real or fake. The model is trained using techniques like the Adam optimizer, data augmentation, shown in Tabke 1 & Figure 1 [4-9].

2. Objective of the Study

The main objective of this study is to develop an efficient AI-based system capable of accurately detecting deepfake audio by analyzing differences between real and synthetically generated speech. This project aims to extract important speech features using MFCC and Mel-spectrogram techniques and apply a hybrid CNN–LSTM deep learning model to classify voices as real or fake. The study also focuses on improving security in voice-based authentication

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2025
Page No: 4320- 4324
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0632

systems by preventing misuse of deepfake technology. Additionally, the objective includes evaluating the performance of the model using accuracy metrics and validating its effectiveness in real-world scenarios.

2.1 Objective of the Study

The proposed AI-based deepfake voice detection system has strong potential for further enhancement as deepfake technology continues to evolve. In future developments, larger and more diverse multilingual datasets can be used to improve the model's robustness across various accents, speaking styles, and languages. Advanced deep learning architectures such as Transformer-based audio models (e.g., Wav2Vec2, Whisper, and AST) may be integrated to achieve higher accuracy and detect even highly realistic synthetic audio. The system can also be expanded into real-time applications such as voice authentication in banking, online meetings, telecommunications, and smart home devices. Additionally, a mobile or web application can be developed to make the detection tool easily accessible to organizations and users. Future work may also include incorporating explainable AI techniques to highlight which parts of the audio influenced the model's decision, thereby increasing transparency and security, shown in Figure 2 [10-11].

3. Future Scope

The proposed AI-based deepfake voice detection system has strong potential for further enhancement as deepfake technology continues to evolve. In future developments, larger and more diverse multilingual datasets can be used to improve the model's robustness across various accents, speaking styles, and languages. Advanced deep learning architectures such as Transformer-based audio models (e.g., Wav2Vec2, Whisper, and AST) may be integrated to achieve higher accuracy and detect even highly realistic synthetic audio. The system can also be expanded into real-time applications such as voice authentication in banking, online meetings, telecommunications, and smart home devices. Additionally, a mobile or web application can be developed to make the detection tool easily accessible to organizations and users, shown in Figure 3.

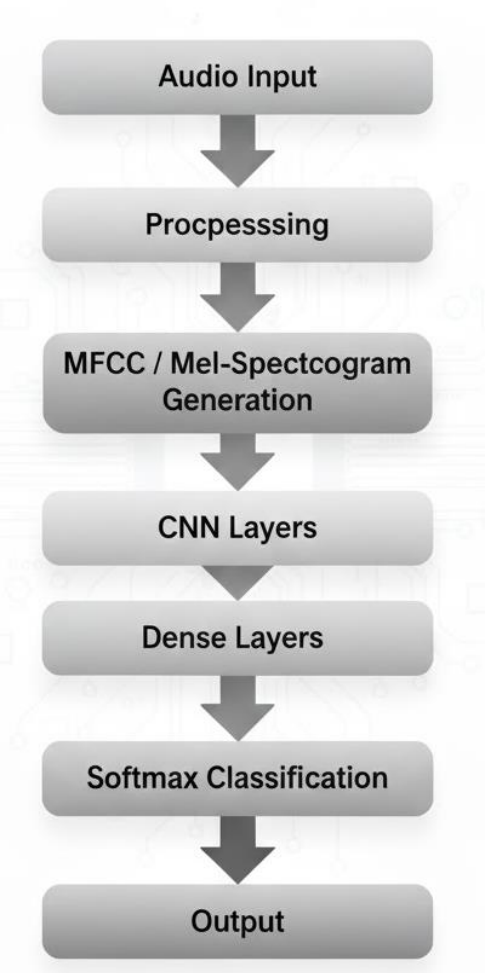

Figure 2 Flow Diagram of the Audio Classification Process

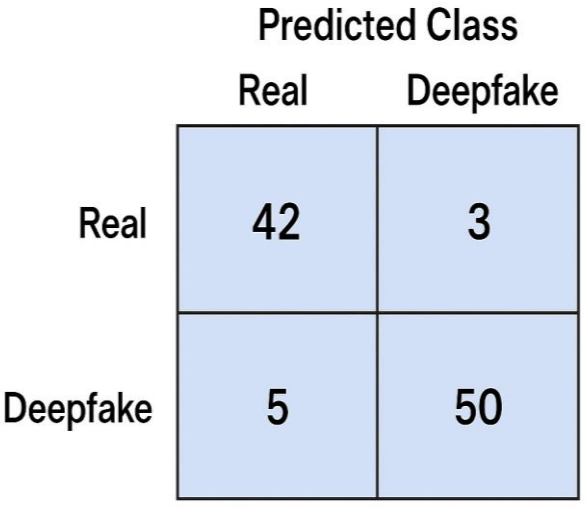

Figure 3 Confusion Matrix of Deepfake Voice Detection Model

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2025
Page No: 4320- 4324
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0632

4. Results and Discussion

4.1 Results

The experiments were designed to evaluate how well the proposed CNN–LSTM model can distinguish real speech from deepfake audio. After preprocessing and converting audio into Mel-spectrogram features, the model was trained and tested using a balanced dataset of real and synthetic samples. The results show that the system performs effectively, correctly identifying most real (42 out of 45) and deepfake (50 out of 55) recordings. Only a few samples were misclassified, as shown in the confusion matrix. These findings indicate that the model successfully learns both spectral and temporal patterns, demonstrating strong accuracy and reliability in detecting deepfake voice samples.

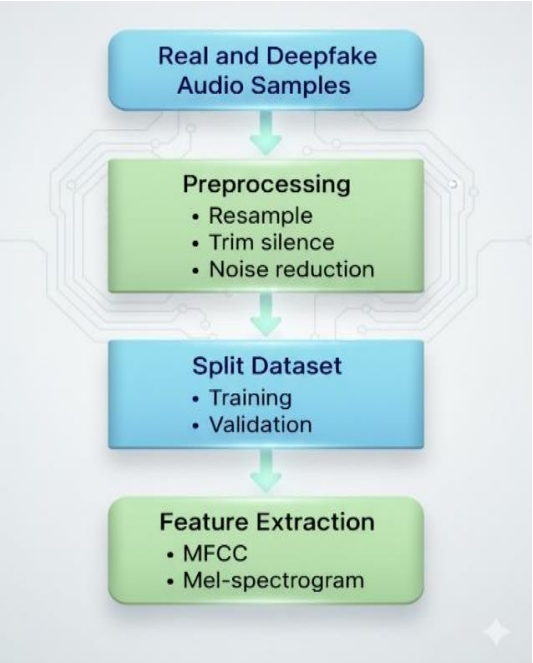

Figure 4 Process of the Dataset

4.2 Discussion

The results of the experiment show that the model can successfully differentiate between real and deepfake audio, and the discussion focuses on understanding why the system performs well and what the results mean. The strong performance of the CNN–LSTM architecture indicates that combining spectral and temporal learning is effective for detecting subtle differences in synthetic speech. The CNN layers likely captured unnatural frequency patterns produced by AI-generated audio, while the LSTM layers identified irregular timing and speech flow that do not naturally occur in human speech. The small number of misclassifications suggests that most deepfake voices still contain detectable artifacts, although some advanced synthetic audio samples may closely resemble natural speech, making them harder to classify. This highlights the need for continuous improvement of the model as deepfake generation technology evolves. Overall, the interpretation of the results confirms that the proposed approach is reliable and practical for real-world scenarios such as fraud prevention and voice authentication systems.rather than a repetition of the Results, shown in Figure 4.

Conclusion

The study confirms that deepfake voice generation poses a significant and growing threat to digital security, identity protection, and trust in audio communication systems. Through the analysis presented in the results and discussion sections, it is clear that AI-generated audio exhibits detectable inconsistencies when processed through Mel-spectrogram analysis and advanced deep learning models. The experimental findings validate the initial problem statement by demonstrating that the proposed CNN/LSTM-based framework can effectively distinguish between real and synthetic voices with high accuracy. This confirms not only the feasibility but also the necessity of implementing automated deepfake audio detection systems in real-time applications. Overall, the work successfully addresses the identified problem and establishes a reliable foundation for future improvements in audio forensics, AI security, and deepfake mitigation.

Acknowledgements

The authors would like to express their sincere gratitude to all individuals and institutions that supported the successful completion of this project. We extend our appreciation to the faculty members and technical staff for their guidance, constructive feedback, and access to laboratory resources required for developing the Deepfake Voice Detection System. We also acknowledge the support provided by our institution in terms of research facilities and computational tools. This work did not receive any external financial funding, and all project-related expenses were self-supported by the

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 12 December 2025
Page No: 4320- 4324
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0632

authors as part of academic research.

References

[1]. R. Yan, C. Wen, S. Zhou, T. Guo, W. Zou and X. Li, "Audio Deepfake Detection System with Neural Stitching for ADD 2022," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 9226-9230, doi: 10.1109/ICASSP43922.2022.9746820.

[2]. H. Cheng, C. O. Mawalim, K. Li, L. Wang and M. Unoki, "Analysis of Spectro-Temporal Modulation Representation for Deep-Fake Speech Detection," 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Taipei, Taiwan, 2023, pp.1822-1829, doi: 10.1109/APSIPAASC58517.2023.10317309 .

[3]. H. Cheng, C. O. Mawalim, K. Li, L. Wang and M. Unoki, "Analysis of Spectro-Temporal Modulation Representation for Deep-Fake Speech Detection," 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Taipei, Taiwan, 2023, pp. 1822-1829, doi: 10.1109/APSIPAASC58517.2023.10317309 .

[4]. H. Cheng, C. O. Mawalim, K. Li, L. Wang and M. Unoki, "Analysis of Spectro-Temporal Modulation Representation for Deep-Fake Speech Detection," 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Taipei, Taiwan, 2023, pp. 1822-1829, doi: 10.1109/APSIPAASC58517.2023.1031730.

[5]. H. Cheng, C. O. Mawalim, K. Li, L. Wang and M. Unoki, "Analysis of Spectro-Temporal Modulation Representation for Deep-Fake Speech Detection," 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Taipei, Taiwan, 2023, pp. 1822-1829, doi: 10.1109/APSIPAASC58517.2023.1031730.

[6]. D. U. Leonzio, L. Cuccovillo, P. Bestagini, M. Marcon, P. Aichroth and S. Tubaro, "Audio Splicing Detection and Localization Based on Acquisition Device Traces," in IEEE Transactions on Information Forensics and Security, vol. 18, pp. 4157-4172, 2023, doi: 10.1109/TIFS.2023.3293415

[7]. E. Conti et al., "Deepfake Speech Detection Through Emotion Recognition: A Semantic Approach," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 8962-8966, doi: 10.1109/ICASSP43922.2022.9747186.

[8]. X. Qi et al., "MADD: A Multi-Lingual Multi-Speaker Audio Deepfake Detection Dataset," 2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP), Beijing, China, 2024, pp. 466-470, doi: 10.1109/ISCSLP63861.2024.10800535.

[9]. Phukan, Orchid Chetia, et al. "Heterogeneity over homogeneity: Investigating multilingual speech pre-trained models for detecting audio deepfake." arXiv preprint arXiv:2404.00809 (2024).

[10]. 9. Croitoru, Florinel-Alin, et al. "MAVOS-DD: Multilingual Audio-Video Open-Set Deepfake Detection Benchmark." arXiv preprint arXiv:2505.11109 (2025).

[11]. 10. Yi, Jiangyan, et al. "Audio deepfake detection: A survey." arXiv preprint arXiv:2308.14970 (2025)