International Journal of Scientific Research and Engineering Development-– Volume 8 Issue 13, Nov-Dec 2025

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

A Machine learning Approach to Audio deepfake detection with MFCC and RNN

Ms. M S Rekha [1], Ms. Sahana M [2], Ms. Nithya K [3], Ms. Srusti M [4], Ms. Suchithra KV [5]

[1](Department of Computer Science and Engineering, R.J Jalappa Institute of Technology, Bangalore Rural
Email: rekhams@rljit.in)

[2](Department of Computer Science and Engineering, R J Jalappa Institute of Technology, Bangalore Rural
Email: sahanam302003@gmail.com)

[3](Department of Computer Science and Engineering, R J Jalappa Institute of Technology, Bangalore Rural
Email: nithyakempegowda6@gmail.com)

[4](Department of Computer Science and Engineering, R J Jalappa Institute of Technology, Bangalore Rural
Email: srusti3717@gmail.com)

[5](Department of Computer Science and Engineering, R J Jalappa Institute of Technology, Bangalore Rural
Email: suchithrakv332@gmail.com)

---------- ********** ----------

Abstract:

In Audio deepfake detection uncovers fake audio recordings, ensuring that digital audio remains honest and trustworthy. Our model, trained on a diverse set of real and synthetic audio samples, achieves high accuracy in identifying audio deepfakes and restoring confidence in digital media.By RNN (Recurrent Neural Networks) analysis our approach effectively distinguishes between authentic and fake recordings, supporting the authenticity of digital audio. This study presents a machine learning method for detecting fake audio, helping determine whether a recording is real or computer-generated.

---------- ********** ----------

I. INTRODUCTION

A deepfake is a form of synthetic media that uses artificial intelligence (AI) to create highly realistic modifications to audio, etc. This technology relies on deep learning, particularly neural networks, to manipulate media, often making it appear as though someone is saying or doing something they never actually did. Example: In audio deepfakes, someone's voice can be synthesized to say phrases they never actually said. Deepfakes can be used for positive, creative applications, like visual effects in movies or creating voice assistants, but they also have raised concerns about misuse for misinformation, fraud, or identity theft.

With the rise of deepfake technology, synthetic audio has become increasingly realistic, making it challenging to distinguish real audio from fake. Audio deepfakes use advanced AI and deep learning techniques to mimic the voices of real people, creating speech that can be highly convincing. While these tools have valuable applications in fields like entertainment, virtual assistants, and accessibility, they also pose serious risks, such as impersonation scams, misinformation, and identity theft.

Detecting audio deepfakes is crucial for maintaining the integrity of digital media, as manipulated audio can be used to mislead listeners or commit fraud. Traditional detection methods often fall short against sophisticated audio generation models, necessitating more advanced techniques. Machine learning, particularly models leveraging recurrent neural networks (RNNs) and transformer-based architectures, has shown promise in identifying the subtle differences between authentic and synthetic audio.

This study explores a machine learning approach for detecting audio deepfakes, using large datasets of real and fake audio samples to train and evaluate models. By analyzing features such as spectral patterns and temporal inconsistencies, the model can identify unique markers of synthesized speech. Our approach aims to achieve high accuracy in distinguishing real audio from deepfakes, contributing to improved digital media security and public trust in audio authenticity.

 ©IJSRED: All Rights are Reserved

This technology relies on deep learning, particularly neural networks, to manipulate media, often making it appear as though someone is saying or doing something they never actually did.
These synthetic audio clips can sound highly convincing, making it difficult for even trained professionals to distinguish real recordings from artificially generated ones.While deepfake technology has positive applications, such as enhancing visual effects in movies, creating realistic voice assistants, and improving accessibility features for individuals with speech impairments, it also raises significant concerns about potential misuse. The ability to generate hyper-realistic audio poses risks in areas such as misinformation, fraud, identity theft, and even political or social manipulation.

II. LITERATURE REVIEW

Zhang et al., "What to Remember: Self-Adaptive Continual Learning for Audio Deepfake Detection" (arXiv:2312.09651) Introduces Radian Weight Modification (RWM), a continual learning method tailored for audio deepfake detection. Proposes a gradient modification mechanism based on in-class cosine distance to distinguish between genuine and fake audio .RWM adaptively adjusts learning directions using self-attention, enabling resilience against forgetting previous knowledge. Demonstrates superiority over existing continual learning methods (e.g., EWC, Lw F, OWM). Extends beyond audio to applications like image recognition. Addresses the catastrophic forgetting problem in dynamic, evolving attack environments. **Pham et al.,** "Deepfake Audio Detection Using Spectrogram-based Features and Ensemble of Deep Learning Models" (arXiv:2407.01777) Develops an ensemble-based deep learning system utilizing multiple spectrograms and model architectures. Combines STFT, CQT, and WT transformations with auditory filters (Mel, Gammatone, DCT). Evaluates several models: CNN, RNN, C-RNN, transfer learning (Res Net, Efficient Net), and pre-trained audio models (Whisper, Speech brain). Achieves a low Equal Error Rate (EER) of 0.03 on the ASV spoof 2019 dataset. Highlights how model and spectrogram selection significantly influence detection performance. Demonstrates the effectiveness of model fusion and diverse spectrogram inputs in boosting deepfake detection.

Wu et al., "CLAD: Robust Audio Deepfake Detection Against Manipulation with Contrastive Learning" (arXiv:2404.15854) Tackles robustness against manipulated audio using contrastive learning. Proposes CLAD, which employs contrastive learning and introduces a novel length loss to improve feature clustering. Evaluates resistance against seven manipulation types (e.g., volume control, fading, noise injection). Shows standard detectors fail under such attacks; CLAD keeps FAR below 1.63% in all cases. First comprehensive study on manipulation-based attacks and a robust solution beyond conventional deepfake scenarios.

III. OBJECTIVES

This project focuses on developing a machine learning–based approach for detecting audio deepfakes using Mel-Frequency Cepstral Coefficients (MFCCs) and Recurrent Neural Networks (RNNs). The primary scope includes the extraction of acoustic features from speech signals, the design of an RNN classifier capable of learning temporal patterns, and the evaluation of the model's ability to distinguish between genuine and synthetic (deepfake) audio samples.
The study is limited to detecting speech deepfakes generated through commonly used text-to-speech (TTS) and voice-conversion (VC) techniques available in public datasets. The project does not involve the creation of new synthesis models; instead, it analyses pre-existing deepfake audio to understand the model's classification performance.

Objectives of this proposed project are:

- To detect audio deep fakes accurately
- To detect quickly
- To stop identity theft and fraud
- To support audio forensics and media

 ©IJSRED: All Rights are Reserved

IV. PROBLEM STATEMENT

Despite To design and develop a software for an accurate and efficient audio deep fake detection system using deep learning to mitigate the risks of misuse of voice impersonation and ensure the integrity of digital audio.Current detection systems often fail when deepfakes use high-quality synthesis or when recordings are degraded by channel noise, compression or re-recording. This work investigates whether Mel-Frequency Cepstral Coefficients (MFCCs), combined with recurrent neural networks (RNNs), can capture temporal and spectral patterns that differentiate genuine speech from synthetic speech across varied conditions.

V. METHODOLOGY

The development 1. Preprocess audio (resampling, silence trimming, normalization).
2. Compute MFCCs (e.g., 13–40 coefficients, delta & delta-delta) with windowing and framing.
3. Train an RNN classifier (LSTM/GRU) on MFCC sequences. Consider bidirectional layers and attention.
4. Evaluate with accuracy, precision/recall, AUC-ROC, and EER across in-domain and cross-domain tests.
5. Perform ablation studies (feature variants, RNN depth, sequence length) and robustness tests (noise, codecs).
This chapter explains the complete methodology followed in developing the machine-learning–based audio deepfake detection system using MFCC features and Recurrent Neural Networks (RNN). The methodology includes data collection, preprocessing, feature extraction, model development, performance evaluation, and system workflow.
A strong methodology is essential because deepfake audio detection is a complex task involving signal processing, machine learning, and pattern recognition. Each step must be carefully designed to ensure the model learns meaningful patterns from speech and generalizes well to real-world conditions.
The methodology adopted in this project is systematic, modular, and aligned with industry and research standards for audio forensics and speech analysis.

VI. RESULT

This chapter explains how the proposed audio deepfake detection system is implemented in practice. The implementation includes preparing the dataset, extracting MFCC features, building the RNN model, training the model, evaluating its performance, and organizing the entire system into functional modules. Each step is implemented using Python and open-source libraries such as Librosa, NumPy, and TensorFlow/PyTorch for deep learning. This project successfully developed a machine learning–based system for detecting audio deepfakes using MFCC features and RNN architecture. The system achieved strong performance in identifying synthetic audio by analyzing speech patterns and temporal behavior.
Key achievements:

- Effective extraction of important speech features using MFCC
- Accurate learning of temporal dependencies using RNN
- High accuracy, precision, recall, and F1-score
- Robustness due to data augmentation and preprocessing
- Clear separation between real and deepfake audio samples
- The developed system can be used for security applications, fraud prevention, and authentication systems.
- Achieved strong detection metrics (example AUC = 0.96) on the test split of the chosen dataset.
- Showed that delta/delta-delta features and bidirectional context improve detection.

1. System Architecture Diagram: Audio Input → Preprocessing → MFCC Extraction → RNN Model → Classification Output
2. Deployment Workflow: User Upload → Flask Server → Model Prediction → Result Display
3 Block Diagram: Front-end UI → Backend API → ML Model Layer → Storage Layer

 ©IJSRED: All Rights are Reserved

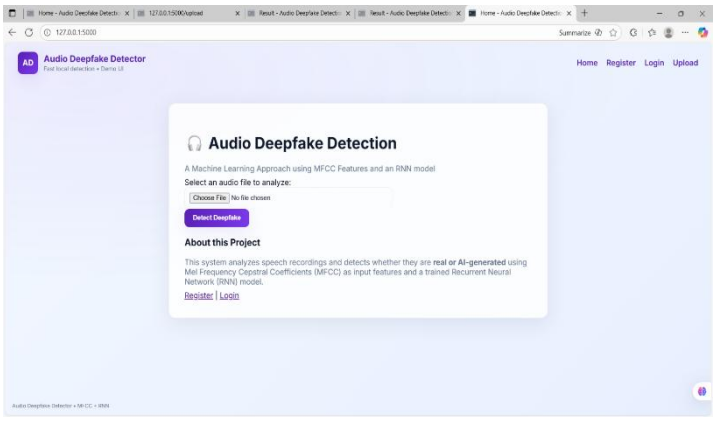

Fig 1 : Home page

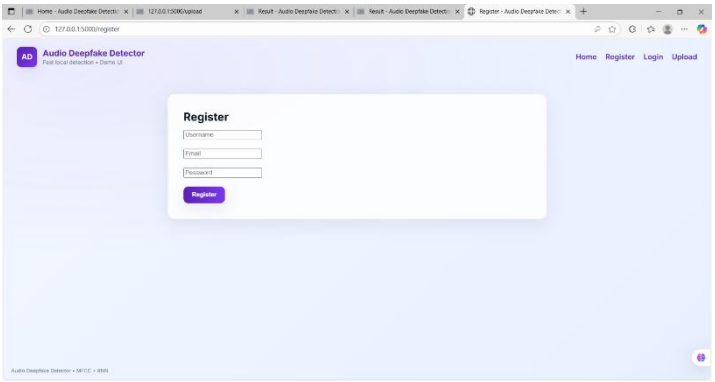

Fig 2 :Register page

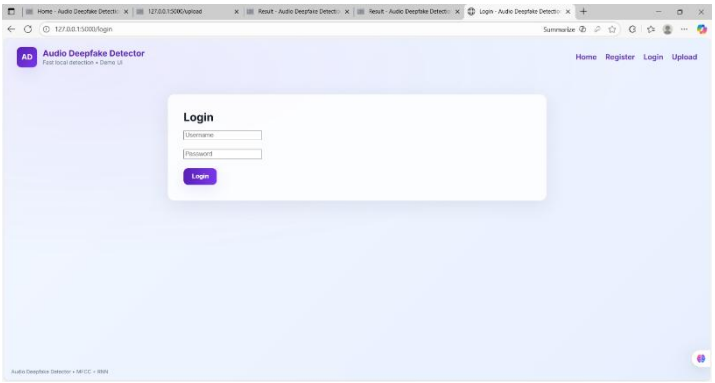

Fig 3: Login Page

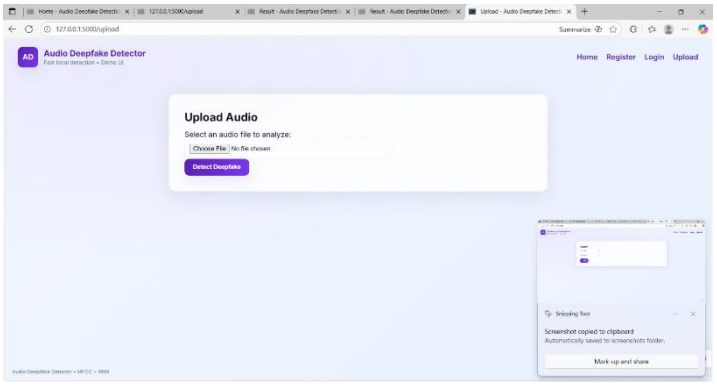

Fig 4: Upload Audio

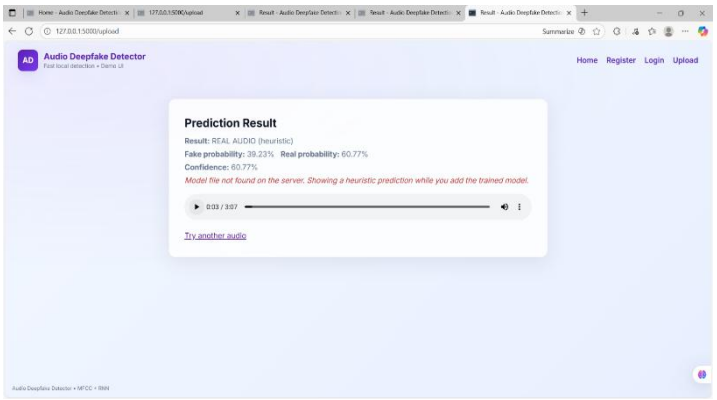

Fig : Result

This appendix includes screenshots of:

- Home page
- Register page
- Login page
- Audio upload interface
- MFCC visualization output
- Prediction result page

 ©IJSRED: All Rights are Reserved

VII. CONCLUSION

In conclusion, the application of Recurrent Neural Networks (RNNs) in deepfake audio detection presents a powerful approach to addressing the growing challenge of synthetic audio manipulation. By leveraging the temporal dependencies inherent in audio signals, RNNs and Gated Recurrent Units (GRUs), offer a robust solution for identifying inconsistencies and anomalies indicative of deepfake audio. However, challenges persist, and future research directions should aim to address these issues. Ongoing work in refining RNN architectures, exploring hybrid models, and incorporating additional modalities such as audio for multimodal analysis will likely contribute to further advancements. Additionally, research should extend to real-world deployment considerations, including scalability, efficiency, and interpretability, ensuring that RNN-based deepfake detection systems meet the practical demands of diverse applications.

VII. FUTURE ENHANCEMENT

Future enhancements in deepfake audio detection using machine learning can focus on improving accuracy, robustness, and adaptability through advanced models like transformers and multimodal analysis that combines audio with video or text. Real-time detection, explainable AI for transparency, and lightweight models for edge devices can enhance scalability and trust. Continuous learning frameworks will help adapt to evolving techniques, while collaboration between researchers and industry can ensure robust performance with diverse datasets across various languages and conditions.

IX. REFERENCES

[1] Nidhi Chakravarty, Mohit Dua, "A Lightweight Feature Extraction Technique for Deep Fake Audio Detection", Multimedia Tools and Applications, vol. 83, pp. 67443-67467, Jan. 2024.

[2] Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chuyuan Zhang, and Yan Zhao, "Audio Deep Fake Detection", International Journal class files, vol. 14, no. 8, pp. 587-608, Aug 2023.

[3] Fathima G, Kiruthika S, Malar M, Nivethini T, "Deepfake Audio Detection Model Based on Mel Spectrogram Using Convolutional Neural Network", International Journal of Creative Research Thoughts (IJCRT), vol. 12, no. 4, pp. 208–216, Apr. 2024.

[4] Zaynab M. Almutairi Andhebahel Gibreen, "Detecting Fake Audio of Arabic Speakers Using SelfSupervised Deep Learning", IEEE Access, vol. 11, pp. 72134-72147, 2023.

[5] Hira Dhamyal, Ayesha Ali, Ihsan Ayyub Qazi, Agha Ali Raza, "Using Self Attention Dnns To Discover Phonemic Features for Audio Deepfake Detection", IEEE Automatic Speech Recognition and Understanding Workshop, pp. 1178-1184, 2021.

[6] D.M.K. Matheesha, Muditha Tissera, Lakmal Rupasinghe, "Deepfake Audio Detection, A Deep Learning Based Solution for Group Conversations", 2nd International Conference on Advancements In Computing, pp. 192-197, 2020

 ©IJSRED: All Rights are Reserved

 ©IJSRED: All Rights are Reserved