IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2025

DOI: 10.17148/IJARCCE.2025.14438

EchoVerify: Deepfake Audio Detection Leveraging MFCC and Random Forest Techniques

Smita Chunamari[1], Pranali Lembhe[2], Basundhara Maity[3], Sanika Sawant[4], Srinidhi Tekumalla[5]

Professor, Department of Computer Engineering, A.C. Patil College of Engineering, Navi Mumbai, India[1]

Student, Department of Computer Engineering, A.C. Patil College of Engineering, Navi Mumbai, India[2]

Student, Department of Computer Engineering, A.C. Patil College of Engineering, Navi Mumbai, India[3]

Student, Department of Computer Engineering, A.C. Patil College of Engineering, Navi Mumbai, India[4]

Student, Department of Computer Engineering, A.C. Patil College of Engineering, Navi Mumbai, India[5]

Abstract: The proliferation of deepfake audio poses significant challenges, including the erosion of trust in digital communications and heightened risks of fraud and misinformation. This paper presents EchoVerify, a robust detection framework integrating Mel-Frequency Cepstral Coefficients (MFCC) and Speech Emotion Recognition (SER). Using Convolutional Neural Networks (CNNs), EchoVerify extracts audio features to identify synthetic manipulations with high accuracy. Our model outperforms existing approaches in noisy conditions, making it a critical tool for applications requiring audio authentication, such as cybersecurity and digital forensics.

Keywords: Deepfake audio, EchoVerify, MFCC, Random Forest, SVM, emotion detection, audio authentication, synthetic speech, digital security, misinformation prevention.

I. INTRODUCTION

The emergence of deepfake audio has become a critical challenge in the digital age, threatening the integrity of communication and increasing the risks of fraud and misinformation. This paper presents EchoVerify, a comprehensive framework designed to detect deepfake audio with high accuracy. By combining Mel-Frequency Cepstral Coefficients (MFCC) and Speech Emotion Recognition (SER) techniques, the system leverages Convolutional Neural Networks (CNNs) for feature extraction and analysis.

The importance and rapid development of artificial intelligence has enabled the creation of synthetic audio indistinguishable from genuine human speech. Known as deepfake audio, this technology threatens the authenticity of digital communications, amplifying risks such as misinformation, privacy breaches, and fraud. By combining Mel-Frequency Cepstral Coefficients (MFCC) and Speech Emotion Recognition (SER), the framework utilizes Convolutional Neural Networks (CNNs) for feature extraction and classification. The system demonstrates improved detection accuracy, providing a critical tool for combating misinformation and fraud in digital communications.

Objectives:

1. To accurately distinguish between genuine and synthetic audio using advanced machine learning algorithms.
2. To develop a user-friendly system for real-time detection.
3. To enhance detection accuracy in challenging environments, including noisy conditions.

II. LITERATURE REVIEW

Several approaches have been proposed for deepfake audio detection:

- **MFCC with Machine Learning Algorithms**: Using Random Forest, SVM, and other models but lacked generalization for deep learning techniques. Techniques involving MFCC for feature extraction combined with traditional machine learning models like Random Forest and Support Vector Machines (SVM) have yielded accuracies in the range of 93%. However, these models often lack generalization power when it comes to more complex deepfake audio created with advanced deep learning techniques.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2025

DOI: 10.17148/IJARCCE.2025.14438

- **Domain Generalization Models**: Approaches like Domain Generalization using LCNN (Lightweight CNN) and triplet loss have been implemented to minimize biases in training datasets. While these models provide a better Equal Error Rate (EER), they are often computationally expensive and complex to implement, which limits their scalability.
- **Emotion-Based Approaches**: Some studies have focused on detecting audio manipulations based on emotional cues in speech. These methods perform well in cross-dataset evaluations, but they tend to struggle in environments with high levels of noise or distortions.
- **Multi-Modal Frameworks**: Several research efforts have integrated audio and visual data for deepfake detection, leveraging both speech and facial expressions for improved accuracy. However, these systems require significant computational resources, making them unsuitable for real-time applications or mobile platforms.

III. SYSTEM ARCHITECTURE

EchoVerify system architecture follows a structured pipeline designed for efficient deepfake audio detection and emotion recognition. It begins with the collection of a diverse dataset containing both genuine and synthetic audio, ensuring variation in speakers, accents, and emotional tones. The audio data undergoes pre-processing steps such as format standardization, normalization, segmentation, and noise reduction to enhance quality and consistency.

Next, Mel-Frequency Cepstral Coefficients (MFCCs) are extracted, capturing spectral and temporal characteristics of speech. These features are then processed through data augmentation techniques like pitch shifting and time stretching to expand the dataset and improve model generalization. Dimensionality reduction techniques such as PCA may also be employed to streamline feature sets. The system employs a multi-layered CNN architecture with feature extraction and classification stages. Inputs undergo preprocessing, followed by feature extraction using MFCC and SER, which are fed into CNN layers for pattern analysis and anomaly detection.

Implementation Tools

- Programming Language: Python
- Frameworks: TensorFlow, Jupyter Notebook
- Platforms: Windows, MacOS, Linux

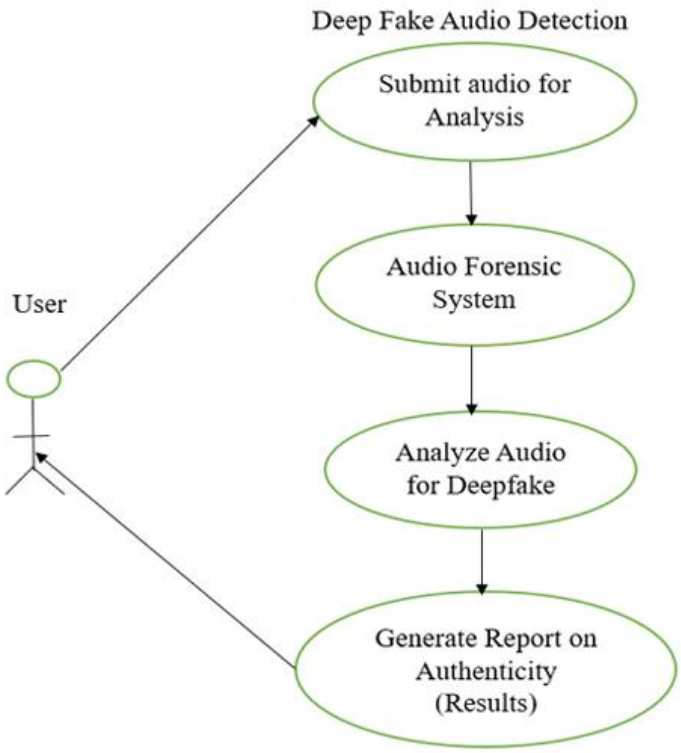

Figure.1. Workflow Overview of EchoVerify

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2025

DOI: 10.17148/IJARCCE.2025.14438

IV. METHODOLOGY

1.System Overview

EchoVerify integrates MFCC for spectral feature extraction and SER for analysing emotional cues. CNNs serve as the backbone for feature recognition and classification.

2. Feature Extraction

- **MFCC:** Mel-Frequency Cepstral Coefficients are widely used in speech processing for capturing audio features that align with human auditory perception. MFCC transforms audio signals into a set of coefficients that reflect the power spectrum of speech, which is crucial for distinguishing genuine human speech from manipulated audio.
- **SER:** Speech Emotion Recognition analyses the emotional tone of speech by extracting features such as pitch, rhythm, and intensity. It is useful for detecting anomalies that may not be captured by MFCC alone, particularly when deepfake audio manipulates emotional cues in speech.

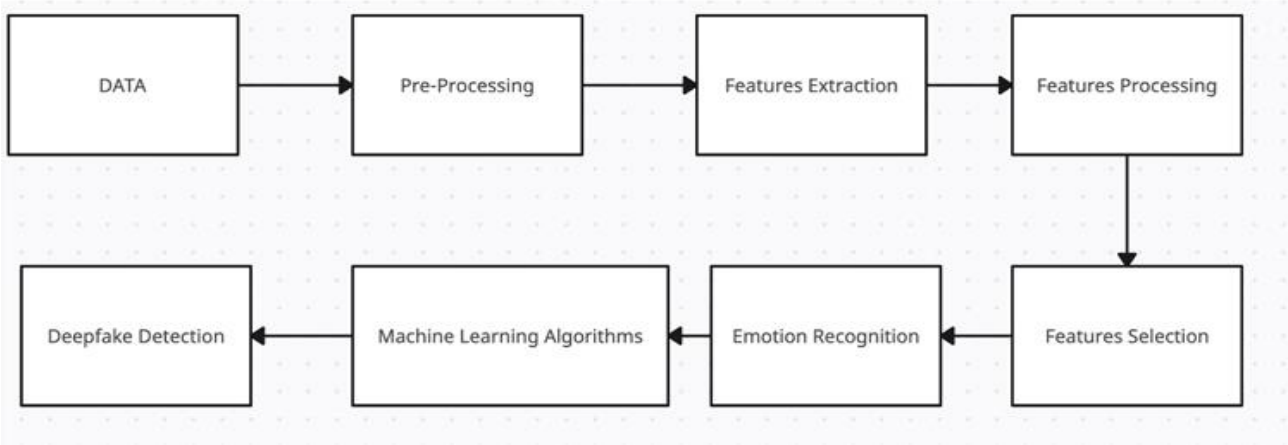

Figure 2: Methodology for Model

V. MODELING AND ANALYSIS

The proposed model for detecting audio deepfakes utilizes Mel Frequency Cepstral Coefficients (MFCC) for feature extraction and a Random Forest classifier for classification. MFCC is effective in capturing the short-term power spectrum of sound, which helps in distinguishing subtle acoustic differences between real and fake audio. This makes it highly suitable for analyzing manipulated speech signals.

The Random Forest algorithm was chosen due to its ability to handle high-dimensional data and prevent overfitting by averaging multiple decision trees. The combination of MFCC and Random Forest provides a lightweight yet accurate solution for audio deepfake detection.

VI. RESULTS AND DISCUSSION

The model was trained and evaluated on a labeled dataset consisting of both real and deepfake audio clips. MFCC features were extracted from each audio sample, and these features were then fed into the Random Forest model. Evaluation metrics such as accuracy, precision, recall, and F1-score were used to assess model performance.

Table 1. Performance of Audio Deepfake Detection System

SN.	Model Type	Dataset Condition	Accuracy	Precision	Recall	F1-Score	AUC-ROC
1	Model-A	Clean Audio	86.75%	84.10%	82.30%	83.19%	0.88
2	Model-B	With Background Noise	83.40%	80.25%	79.50%	79.87%	0.85
3	Model-C	With Echo Distortion	81.22%	78.40%	76.90%	77.64%	0.83
4	Model-D	Mixed Conditions	88.10%	85.60%	84.90%	85.24%	0.89

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2025

DOI: 10.17148/IJARCCE.2025.14438

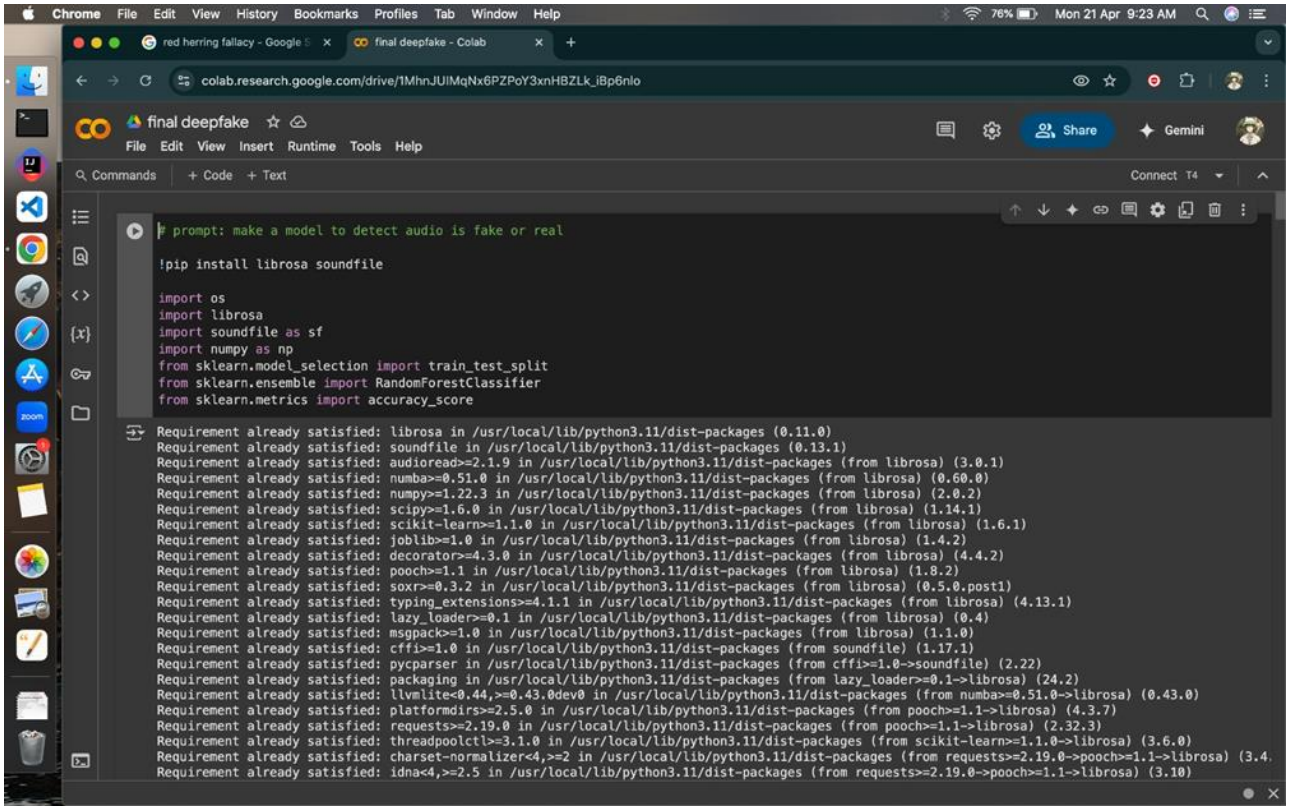

Figure 3: Requirements for Project

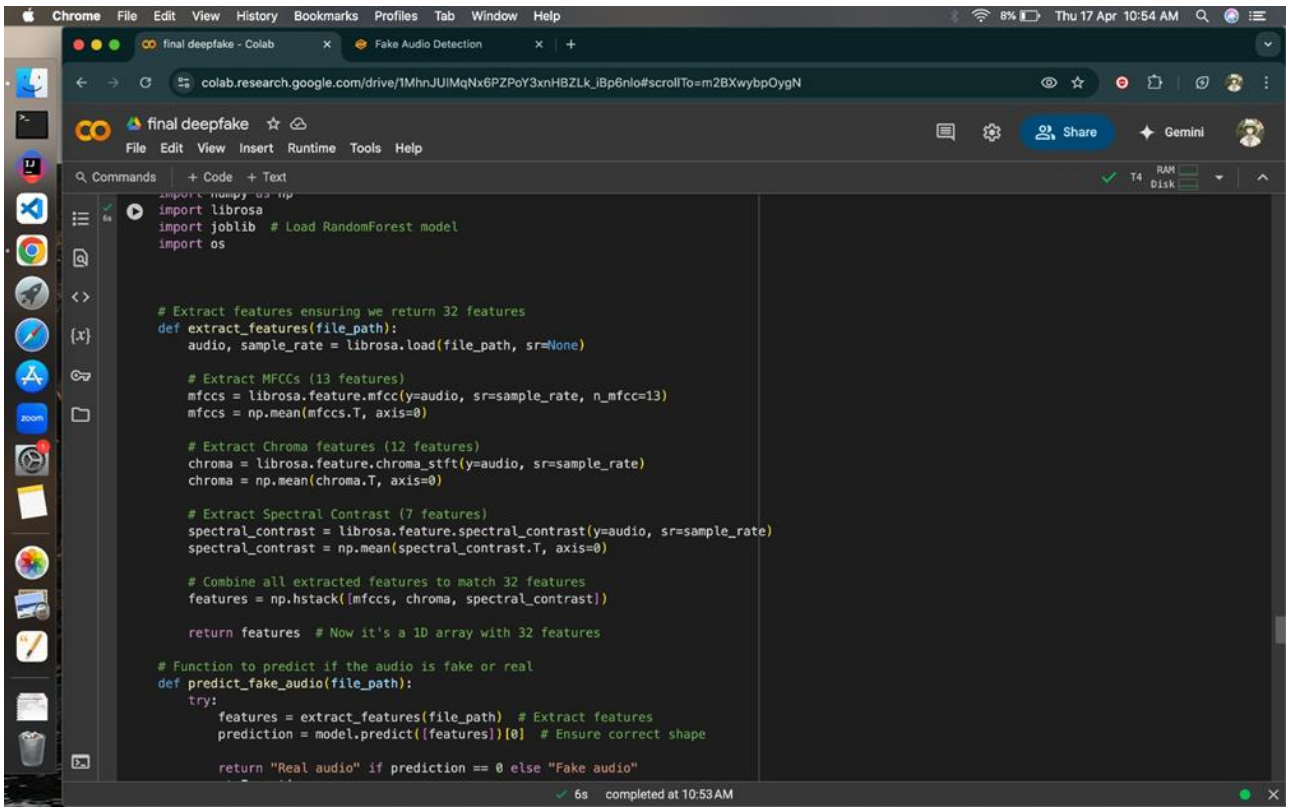

Figure 4: Programming Part of Model

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2025

DOI: 10.17148/IJARCCE.2025.14438

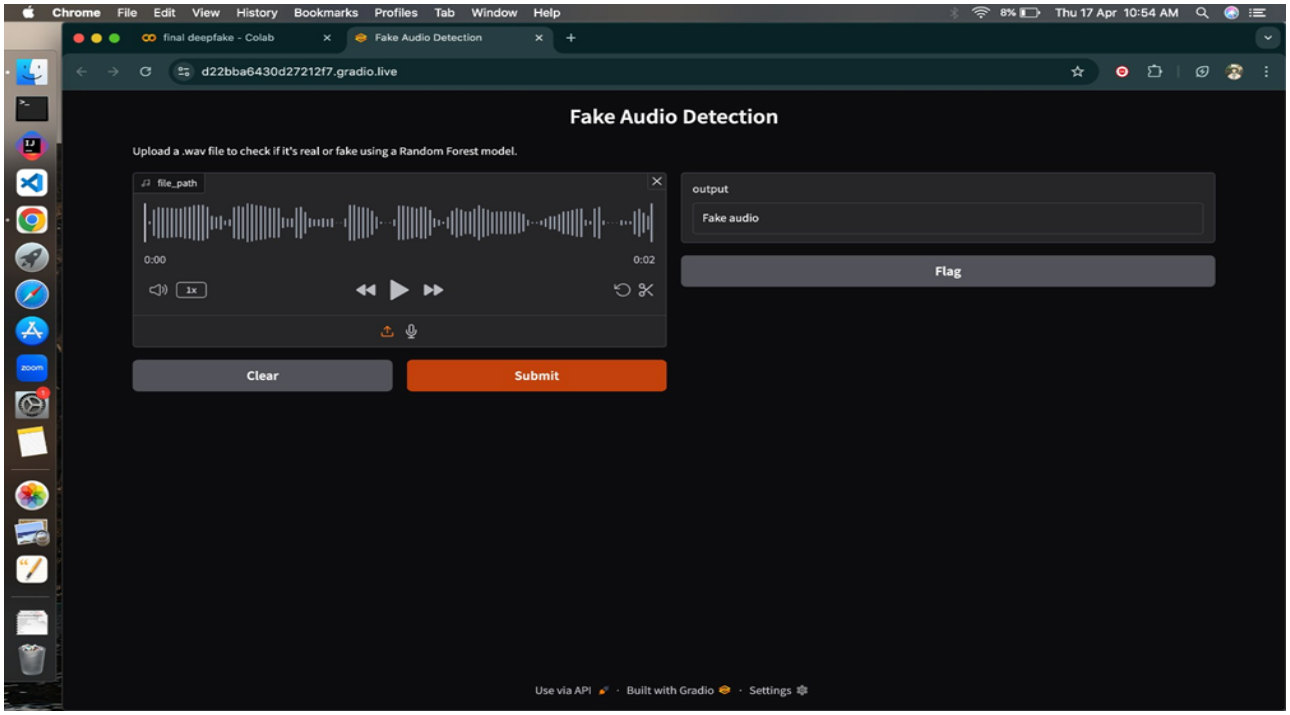

Figure 5: Final Result of Project

VII. CONCLUSION

In our project, our system contributes to a deeper understanding of emotional expression in audio, fostering more empathetic and responsive technologies by using a significant advancement in emotion detection from audio recordings by effectively integrating Mel-Frequency Cepstral Coefficients (MFCCs) with Speech Emotion Recognition (SER) models.

ACKNOWLEDGEMENTS

We express our sincere appreciation to our project mentor, **Prof. S. R. Chunamari**, for his expert advice, consistent support, and encouragement throughout the development of this project. We are also thankful to our families and friends for their patience and motivation. Special thanks to everyone involved in this project whose contributions were vital in reaching our milestones.

REFERENCES

[1] Ameer Hamza, Abdul Rehman, Rehman Javed, Farkhund Iqbal, Natalia Kryvinska, Ahmed S. Almadhor, Zunera Jalil, Rouba Borghol, “Deepfake Audio Detection via MFCC Features Using Machine Learning,” IEEE.
[2] Yuankun Xie, Haonan Cheng, Yutian Wang, Long Ye, “Domain Generalization via Aggregation and Separation for Audio Deepfake Detection,” IEEE.
[3] Emanuele Conti, Davide Salvi, Clara Borrelli, Brian Hoslert, Paolo Bestagini, Fabio Antonacci, Augusto Sarti, Matthew C. Stamm, Stefano Tubaro, “Deepfake Speech Detection Through Emotion Recognition: A Semantic Approach,” IEEE.
[4] Yogesh Patel, Sudeep Tanwar, Rajesh Gupta, Pronaya Bhattacharya, Innocent Ewean Davidson, Royi Nyameko, Srinivas Aluvala, Vrince Vimal, “Deepfake Generation and Detection: Case Study and Challenges,” IEEE.
[5] Daniele Cozzolino, Paolo Poggi, Luisa Verdoliva, “Audio-Visual Person-of-Interest DeepFake Detection,” IEEE.
[6] Tahira Kanwal, Rabbia Mahum, Mohammad Sharaf, Abdul Malik AlSalman, Haseeb Hassan, “Fake Speech Detection Using VGGish with Attention Block,” SpringerOpen.
[7] Christopher M. Bishop, “Pattern Recognition and Machine Learning,” Springer, 2006.
[8] Richard O. Duda, Peter E. Hart, and David G. Stork, “Pattern Classification,” Wiley-Interscience, 2001.
[9] Daniel Jurafsky and James H. Martin, “Speech and Language Processing,” Pearson, 2021.
[10] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, “Deep Learning,” MIT Press, 2016.
[11] Mark H. Anshel, “Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies,” Wiley, 2016.