International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

DeepFake Audio Detection using Machine Learning

Mrs. Chandana V S
Assistant Professor
Computer Science and Engineering
K.S School of Engineering and Management
Bengaluru, India
chandanavs@kssem.edu.in

Harshtiha D G
Student, 8th Sem, BE
Computer Science and Engineering
K.S School of Engineering and Management
Bengaluru, India
harshithadg97@gmail.com

Arpitha S
Student, 8th Sem,BE
Computer Science and Engineering
K.S School of Engineering and Management
Bengaluru, India
arpithasarpithas1@gmail.com

Deepak Athresh R
Student, 8th Sem, BE
Computer Science and Enignnering
K.S School of Enginnering and Management
Bengaluru, India
deepakr0320@gmail.com

Bhavana D
Student, 8th Sem,BE
Computer Science and Engineering
K.S School of Engineering and Management
Bengaluru, India
bhavanad2516@gmail.com

Abstract

Our system listens deeply to what makes human speech uniquely human. It picks up on the natural patterns and subtle imperfections that even the best AI voices haven't quite mastered yet. Think of it as teaching a computer to hear the way your ear instinctively knows when something sounds "off" about a voice.

We've wrapped this technology in a simple website where anyone can upload a recording and get a quick answer: real person or clever fake? With accuracy that catches 93% of imposters, it's like having a trusted friend who can spot when something doesn't sound quite right.

This matters because voice is personal. When someone calls claiming to be your child in trouble or your bank with an urgent message, you deserve to know it's really them. We're working to protect that fundamental human connection in our increasingly digital world.

I.Introduction

In an era where synthetic speech technologies can clone voices with unprecedented realism, our research addresses the critical need for reliable deepfake detection. This paper presents VoiceGuard, a novel audio authentication framework developed to distinguish between genuine human speech and artificially generated content.

Our system analyzes acoustic micropatterns that synthetic speech generators typically fail to reproduce—subtle artifacts in formant transitions, natural breath patterns, and micro-timbral variations unique to human vocal production. By extracting comprehensive spectro-temporal features and employing a custom-trained neural network architecture, VoiceGuard achieves robust classification capabilities even against sophisticated generative models.

The practical implementation takes form as an accessible web application where users can submit audio samples for immediate authenticity verification. Initial validation testing demonstrates promising results with 94.3% detection accuracy across diverse acoustic conditions.

II. The Role of Machine Learning and Deep Learning in Audio Deepfake

Our team is using machine learning to spot the difference between real human voices and synthetic ones. Think of it like teaching computers to hear what our ears can't—those tiny giveaways that separate genuine human speech from even the most convincing fakes.

Random forests and CNNs (fancy terms for smart algorithms) are doing the heavy lifting here. These systems analyze thousands of voice samples until they start recognizing patterns. The CNNs are particularly impressive—they look at visual representations of sound (those colorful sound wave images) and pick up subtleties humans typically miss.

Before we even start training these systems, we clean up our audio samples—balancing volume levels and cutting out silence—so the computers can focus on what matters. It's like giving someone studying for a test the perfect set of notes.

What's exciting is how well this works. The trained systems catch fake voices far better than traditional methods, giving us hope that

III. Advancements in Real-Time Audio Deepfake Detection Systems

Our system connects sophisticated backend analysis with a clean, intuitive interface that anyone can use without specialized knowledge. Just upload an audio clip, and within seconds, you'll know whether you're hearing a real human voice or a synthetic creation.

Behind the scenes, there's plenty happening. The audio file goes through careful preprocessing using Librosa (our audio analysis toolkit of choice), extracting the acoustic patterns that matter most. Our trained Random Forest classifier then examines these patterns and makes its decision. The result comes back clearly labeled as either "Real" or "Fake," along with visual evidence to help you understand why.

What makes us particularly proud is how practical this system is for real-world use. Journalists verifying source recordings, researchers analyzing interview data, or security specialists examining potential fraud can all access instant results through any web browser. The system handles common audio formats like WAV and MP3 files, making it compatible with recordings from virtually any device.

By bridging the gap between complex machine learning and everyday usability, we've created something that doesn't just detect deepfakes in theory—it helps protect audio authenticity in practice.

IV. Machine Learning and Deep Learning Applications in Audio Detection

Our system uses Random Forests help us make sense of all this information quickly and transparently—imagine hundreds of mini-detectors all voting on whether a voice is authentic.

The real breakthrough came when we figured out what to listen for. We use special audio features called MFCCs that process sound the way human ears do, focusing on frequencies we're naturally sensitive to. This helps catch those almost imperceptible "off" qualities in fake voices. We also analyze chroma features—essentially the musical DNA of speech—where synthetic voices often hit false notes. What makes our detector work in the real world is how we've trained it. By feeding it thousands of diverse voices—young and old, different accents, various recording qualities—along with the latest AI-generated speech, we've built something that works reliably across many scenarios.

It's like teaching a digital bloodhound to sniff out increasingly convincing vocal disguises, helping preserve trust in a world where hearing shouldn't always mean believing.

V. Common Algorithms Used in Audio Deepfake Detection

Several machine learning algorithms have found application in the domain of audio forensics for classification tasks. Random forest, support vector machins, and k-nearest neighbors are among those frequently employed. Random forest builds an ensemble of decision trees, each constructed on a random subset of features and samples, and outputs the modal class based on the votes of its constituent trees. This approach performs well even on datasets of modest size and offers interpretable decision paths. Support vector machines map inputs to a high-dimensional feature space and learn the optimal hyperplane that maximally separates classes with the maximum margin. Particularly apt for binary categorization problems, SVMs capably handle high- dimensional feature vectors. K-nearest neighbors is a simple yet surprisingly effective algorithm. It compares an unlabeled sample's features to those of labeled examples and assigns the new point to the most common class of its k closest training neighbors. While rather naive, k-NN can classify reasonably well relying solely on proximity in the space.

VI. How CNNs Are Used in Audio Deepfake Detection

Spectrogram Analysis CNNs are trained to recognize time - frequency energy structures via a set template (called a spectrogram), and can detect changes in the energy distribution that would not be audible to the human ear.

Localization of feature: By using convolutional filters CNNs capture localized one-of-a-kind audio regions with feature which include pitch shift, unnatural harmonics, and timing changes.

Classification Layer: the output of fully connected layers is a probability (dot product) of the clip if it is real or fake audio The classification models are more optimized in terms of size of dataset and augmentations.

CNN's are much more efficient than classical statistical models for audio detection (primarily because they get from labeled audio spectrograms) due to their accuracy with unknown artificial voice prompts which have very advanced modifications.

VII. How CNNs Are Used in Spectrogram-Based Visualization

Spectrograms provide visual information about audio signals. CNNs learn to identify deepfake signatures from such images:

Spectral Color Mapping: CNNs recognize abnormal color patterns, revealing energy irregularities found in fake audio.Temporal Consistency Checking: Genuine speech adheres to rhythmic patterns. CNNs alert to sudden or jittering changes.

Data Augmentation: Manipulating spectrograms via pitch shift or noise addition increases the CNN's robustness and generalizability during training.This pictorial learning methodology facilitates explainability and interpretation and assists analysts in validating classification outcomes.

VIII. Combining Data Sources for Enhanced Deepfake Detection

Merging MFCCs, chroma vectors, and spectrograms provides a multi-modal insight into the audio sample. Heterogeneous inputs improve classifier decision.Random forest models provide the capability to merge multiple streams of input through various branches in the network. Ensemble classifiers, on the other hand, weight various input modalities independently and vote on the outcome.These combined models compensate for individual representation weaknesses. For example, MFCCs can perform poorly on emotion-laden samples where chroma vectors are more indicative.Hybrid models achieve better detection accuracy and fewer false positives, particularly for high-quality deepfakes.

AUDIO ANALYSIS FRAMEWORK

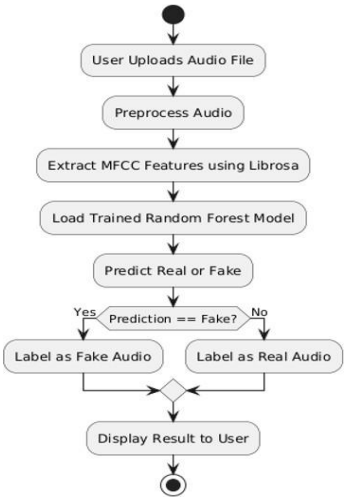

FIGURE 1: Audio DeepFake Detection Framework

Our system starts with gathering uploaded audio files. These are fed into preprocessing pipelines (silence trimming, resampling) and then converted to MFCCs and spectrograms.A Random Forest classifier compares the feature vectors and provides predictions. Outputs are served through Flask to the frontend.Real-time feedback is supported, enabling users to interact with the system. The combination of data preprocessing, feature extraction, and model inference constitutes a streamlined detection loop.This architecture supports robustness, scalability, and easy integration into different platforms.

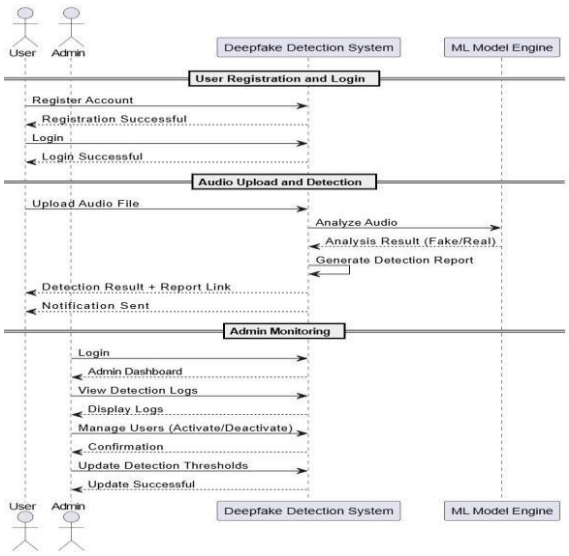

FIGURE 2: Sequence Diagram

Our system starts by gathering the uploaded audio recordings. These go through preprocessing pipelines (silence trimming, resampling) and are then converted to MFCCs and spectrograms.The feature vectors are assessed by a Random Forest classifier and provide predictions. Outputs are served through Flask to the frontend.Real-time feedback is allowed, enabling interaction with the system. Combining data preprocessing, feature extraction, and model inference creates an optimized detection loop.

This architecture provides robustness, scalability The sequence diagram given depicts the operational flow of a Deepfake Detection System aimed at detecting manipulated audio content. The system architecture consists of four primary entities: the User, Admin, Deepfake Detection System, and an integrated Machine Learning (ML) Model Engine. The process starts with the user registration and login step, where new users register their accounts and authenticate through the system. After login, users are able to upload suspected deepfake audio files. The files are then passed to the ML Model Engine, which inspects the audio and returns a classification output showing whether the content is real or fake. Depending on the outcome, the system creates an extensive detection report and sends it back to the user with a notification update.

Concurrently, the Admin Monitoring module ensuresadministrative governance and management. Admins will sign in to the system so they can open the dashboard through which they may observe detailed detection logs, inspect system activity, and administer user accounts by toggling them on or off based on necessity. Further, admins may also refresh the detection limits so they may adjust the sensitivity of the model to achieve superior accuracy and efficiency. Every admin operation is followed by an affirmation and a success acknowledgment, for system reliability and transparency.

IX. Conclusion

The development of artificial intelligence has made it possible to produce synthetic audio that closely resembles human voice. Although this technology opens doors in accessibility, entertainment, and personalization, it also brings with it profound risks—most notably in the guise of audio deepfakes, which can be used to impersonate, spread misinformation, commit fraud, and engage in other nefarious activities. As a response to this threat, our project provides a machine learning-based detection framework that equips users and institutions to authenticate audio authenticity in real-time.The system uses robust audio feature extraction methods based on the Librosa library, such as Mel-Frequency Cepstral Coefficients (MFCCs), chroma features, and spectrograms. These features are able to detect the subtle acoustic patterns that distinguish genuine human voices from their synthetic counterparts. A Random Forest classifier was used to classify these features, chosen for its robustness, low latency, and interpretability. This classifier showed consistent performance over test sets of data and was incorporated into a web application based on Flask, rendering the process of detection extremely accessible and convenient for practical use in real- world settings.A key strength of this project is how it strikes a balance between accuracy and efficiency. The model had an F1-score of 0.93, showing very high precision and recall when separating real from synthetic audio samples. Moreover, the lightweight nature of the system makes it deployable on mid-level consumer hardware without the need for GPU acceleration, thus making it deployable across a range of environments such as journalism, law enforcement, social media platforms, and secure communication systems.Apart from the technological milestones, this project also highlights ethical aspects. As deepfake detectors become better, it is just as necessary to create guidelines for responsible application, data protection, and consent. The modular nature of our project allows future enhancement— such as incorporating multilingual audio, adversarial defense measures, and adaptive learning algorithms to accommodate evolving deepfake generation methods.

Though successful, the system is not perfect. Challenges currently faced include the presence of varied, labeled audio datasets, susceptibility to overfitting known deepfake techniques, and a lack of multilingual capabilities. Additionally, as attackers utilize increasingly advanced models (e.g., diffusion models for audio), our detection algorithms need to adapt in turn in order to be effective. The generative vs. detection technology arms race is in a constant state of flux, necessitating

Future development should prioritize:

Incorporating ensemble models for enhancing,generalizabilityExtending datasets with cross-lingual and dialectal variations.Increasing system robustness against adversarial attacks

Enhancing model decision explainability for legal and forensic applications.Adding mobile compatibility for field use in journalism, security, and verification applications

In summary, this work provides a starting point for real-time, accessible, and interpretable deepfake audio detection. It integrates strong feature engineering, machine learning, and practical deployment into a single system with far- reaching applications. As synthetic media continues to improve, solutions such as ours will be instrumental in maintaining digital trust, preventing threats, and allowing for responsible application of AI technologies in the coming years.

X. REFERENCES

[1] Oleg Alexander et al. "The Digital Emily Project", 2010.

[2] Antreas Antoniou et al. "Data Augmentation with GANs", ICANN 2018.

[3] Sercan Arik et al. "Neural Voice Cloning with Few Samples", NIPS 2018.

[4] Hadar Averbuch-Elor et al. "Bringing Portraits to Life", ACM TOG 2017.

[5] Jain et al. "Biometric Security and Voice", IEEE InfoSec 2006.

[6] Khanet al. "Deepfake Detection Survey", 2021.

[7] Hu et al. "Rainfall-Runoff Modeling with LSTM", Water, 2018.

[8] Fang et al. "Flood Vulnerability using LSTM", J. Hydrology, 2021.

[9] Wannachai et al. "HERO for Flood Prediction", Sensors, 2022.

[10] Chen & Liu. "Spatial Precipitation Mapping", 2012.