iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** I **Month of publication:** January 2025

DOI: https://doi.org/10.22214/ijraset.2025.66025

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue I Jan 2025- Available at www.ijraset.com

AudioVeritas: A Machine Learning Model to Detect Deepfake Audio

Ganavi M[1], Shashank R R[2], S Varun[3], Sanketh Salanke R[4], Vishal S Navale[5]
[1, 2, 3, 4, 5]*Department of Computer Science and Engineering, Jawaharlal Nehru New College of Engineering*

Abstract: ***With the rapid advancement of deep learning technologies, the creation of synthetic media, particularly deepfake audio, has become increasingly prevalent. Deepfake audio convincingly replicates human voices, presenting both innovative opportunities and significant risks, including misinformation and fraud. Detecting such audio is challenging due to the subtle differences from real speech, often imperceptible to human hearing. This paper introduces a machine learning-based framework for detecting and classifying audio as "REAL" or "DEEPFAKE." Leveraging signal processing techniques, including Mel-Frequency Cepstral Coefficients (MFCCs), Chroma Features, and Spectral Properties, our approach identifies patterns distinguishing real audio from synthetic. The proposed system addresses ethical and security concerns surrounding synthetic audio misuse while contributing to advancements in deepfake detection research.***

Keywords: Audio Authenticity Verification, Deepfake Detection, MFCC and Chroma Features, Signal Processing, Digital Forensics.

I. INTRODUCTION

With the rise of deep learning technologies, synthetic media creation has become increasingly accessible and sophisticated. Among these, deepfake audio has emerged as a significant innovation, allowing machines to replicate human voices with impressive accuracy. While this technology finds applications in areas like personalized voice assistants, audiobooks, and entertainment, its misuse has led to growing concerns. Deepfake audio can be weaponized for malicious purposes, such as impersonating individuals to spread misinformation, commit fraud, or manipulate public opinion. The detection of deepfake audio presents a unique challenge because synthetic voices can closely mimic natural patterns of speech, including tone, pitch, and rhythm. Subtle differences between real and fake audio often go unnoticed by the human ear, necessitating the development of automated tools to analyze and classify audio files effectively. This work leverages machine learning and signal processing techniques to detect and classify audio as "REAL" or "DEEPFAKE." The system is built using a pipeline of feature extraction and classification, wherein audio characteristics are analyzed to capture underlying patterns. By extracting features such as MFCCs, Chroma Features, and Spectral Properties, the proposed system identifies distinguishing traits between authentic and synthetic audio, enabling accurate classification.

This automated framework not only highlights the potential of artificial intelligence in combating deepfake misuse but also demonstrates the power of machine learning in solving real-world security and ethical challenges.

II. LITERATURE SURVEY

In [1] the authors introduce WaveFake, a novel dataset designed for audio deepfake detection. They address the need for a comprehensive dataset by generating audio samples using six SOTA architectures across two languages. They propose a thorough frequency analysis to identify subtle differences between architectures. They implement and evaluate three baseline classifiers, including GMMs and RawNet2, to provide benchmarks for future research. The dataset allows one-to-one comparisons across architectures, enabling robust model training and evaluation. The classifiers offer a baseline to measure the performance of new detection methods. However, the dataset is limited to certain architectures and languages, potentially limiting its real-world applicability. Additionally, it might not fully represent real-world conditions due to a lack of diverse speakers and scenarios.

In [2] the researchers propose a method to detect deepfake voices using explainable deep learning techniques. They use simple CNN and LSTM-based architectures to maintain interpretability. The study utilizes spectrogram-based features and datasets like ASVspoof and LJSpeech. Deepfake audio samples are generated using TTS and voice conversion methods. XAI techniques like LRP, Deep Taylor, and Integrated Gradients are applied to highlight influential regions of input audio. The results reveal key characteristics of deepfake speech, such as lower pitch variance and rhythmic monotony. The study reconstructs attribution scores into audio format for intuitive interpretation. By focusing on explainability, the authors aim to address the "black-box" nature of deepfake detection models. This work contributes to the development of reliable and transparent AI models for audio security and media authenticity verification.

In [3] the authors propose a Fake Voice Detection System that utilizes CNNs and RNNs to differentiate between authentic human voices and AI-generated fake voices. The system employs Tortoise-TTS to generate synthetic voices for training and extracts audio features like MFCCs, spectrograms, and waveform characteristics. The deep learning models are trained to identify patterns indicative of fake voices. The system is adaptable and scalable, demonstrating robust performance even under challenging conditions. However, its reliance on specific audio features may limit its ability to detect highly sophisticated deepfake techniques that manipulate audio characteristics not covered by the current feature set.

In [4] the authors provide a comprehensive review of audio deepfake detection, surveying various techniques and methodologies. The survey analyzes different types of deepfake audio, popular datasets, and feature extraction methods. Both pipeline-based and end-to-end detection approaches are evaluated. The paper highlights state-of-the-art datasets and classifiers, providing a systematic comparison. The authors identify critical gaps in the field, such as the lack of large-scale, real-world datasets and the need for model interpretability. However, the paper has limitations in discussing model interpretability and the generalization ability of current models due to limited datasets.

In [5] the author provides an overview of librosa, a Python library by Brian McFee et al., for audio and music signal processing. It addresses tasks like spectrogram calculation, pitch operations, and onset detection, offering a user-friendly API. Integration with libraries like NumPy and SciPy ensures efficient and flexible audio analysis. The library proposes a standardized approach to audio processing, supporting reproducibility and accessibility for MIR researchers, including those transitioning from MATLAB. Its strengths include ease of use and compatibility within the Python ecosystem. However, librosa has limitations, such as reduced customization due to default parameters and storage challenges from its disk-based caching in large datasets. Despite this, it remains a powerful and versatile tool for MIR and audio analysis. Users should be mindful of its storage and customization challenges in specialized applications.

In [6] the researcher explores audio deepfake detection using Mel-frequency cepstral coefficients (MFCCs) and models like SVM and VGG-16. The "Fake-or-Real" dataset, segmented by audio length and bit rate, was used for evaluation. VGG-16 showed strong performance through transfer learning, achieving high detection accuracy. The study highlights the effectiveness of combining MFCCs with advanced classifiers and provides a foundation for future research. However, VGG-16's high computational demands and challenges with diverse datasets like "for-original" indicate a need for optimization. Despite these issues, the research demonstrates the potential of deep learning in synthetic audio detection.

III. OBJECTIVES

1) Develop an audio processing pipeline to normalize and analyze audio signals.
2) Extract meaningful features, including MFCCs, Chroma Features, Spectral Centroid, Bandwidth, Zero-Crossing Rate and train a Support Vector Machine (SVM) to classify audio as "REAL" or "DEEPFAKE".
3) Evaluate model performance using metrics such as accuracy, precision, and F1-score.
4) Implement a user-friendly interface for testing and training.

IV. METHODOLOGY

This proposed work focuses on audio deepfake detection by leveraging machine learning techniques, specifically using a Support Vector Machine (SVM) classifier. It begins with essential preprocessing, including audio normalization, to standardize signal amplitude and mitigate bias caused by varying audio volumes. This extracts a rich set of audio features such as MFCCs (Mel-Frequency Cepstral Coefficients), chroma features, spectral centroid, spectral bandwidth, and zero-crossing rate, which collectively represent the timbral, harmonic, and spectral characteristics of audio files. These features are crucial for distinguishing between real and synthetic (deepfake) audio samples.

The dataset is structured into labeled categories—original for real audio and deepfake for synthetic audio—allowing the system to assign binary labels (1 for real and 0 for fake) during training. Using these labeled datasets, the features are extracted and aggregated into a feature matrix. This data is then split into training and testing subsets, with 80% of the data used for model training and 20% for evaluation. The SVM classifier, known for its effectiveness in binary classification tasks, is trained using a linear kernel. The model's performance is evaluated through metrics like precision, recall, F1-score, and accuracy, ensuring robust classification capabilities. A key strength of the model lies in its modular design and reproducibility. The trained model is saved for reuse, enabling seamless testing on new audio samples. Additionally, user interaction is incorporated, allowing users to train the model with their dataset or directly test a pre-trained model on new audio files. The model employs tools like tqdm for progress tracking, enhancing user experience during feature extraction and training.

While the work effectively addresses the core task of detecting audio deepfakes, several areas for improvement have been identified. These include expanding support for additional audio formats, addressing potential dataset imbalances through techniques like oversampling or class weighting, and enhancing feature extraction by integrating advanced features such as Mel-spectrograms. Furthermore, implementing cross-validation would ensure better model generalization, and adding support for real-time audio analysis would broaden the applications in fields like cybersecurity, fraud detection, and media verification. Overall, the proposed work represents a comprehensive approach to audio deepfake detection while offering avenues for future enhancements.

V. SYSTEM DESIGN

The system is designed to detect audio deepfakes through a machine learning pipeline that involves preprocessing, feature extraction, model training, and user interaction as presented in Figure 1. It focuses on delivering accurate classifications while ensuring flexibility and ease of reuse. The system also supports model updates and loading for future improvements.

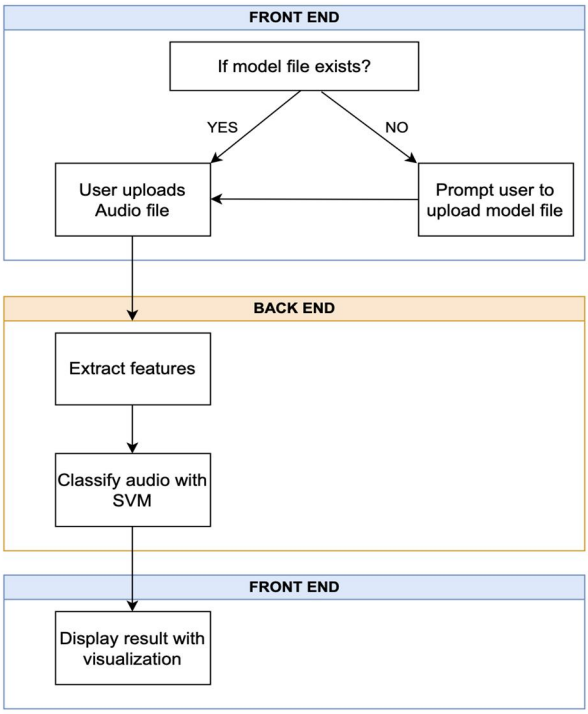


Fig 1: System Architecture

A. *Audio Preprocessing*

In this step, the audio is first normalized to ensure consistent loudness across different audio files. After normalization, meaningful features are extracted from the audio, such as MFCCs (Mel Frequency Cepstral Coefficients), Chroma, Spectral Centroid, Spectral Bandwidth, and Zero-Crossing Rate. These features are important for feeding into a machine learning model, as they capture key characteristics of the audio signal.

B. *Model Training*

The dataset, consisting of labeled REAL and DEEPFAKE audio files, is used to train the model. Features are extracted from these files, and the data is fed into a Support Vector Machine (SVM) model for training. The model is then evaluated using a test dataset to assess its performance, focusing on accuracy and classification metrics. After evaluation, the trained model is saved for future use, allowing it to be loaded and used for predictions on new audio files.

C. *Audio Classification (Prediction)*

When a new audio file is uploaded, the audio is preprocessed by extracting the same features used during training. The extracted features are then fed into the trained model for classification, and the model predicts whether the audio is REAL or DEEPFAKE. The result is displayed to the user, informing them of the classification outcome.

D. Frontend (User Interface)

The user can upload an audio file through a simple interface that allows for easy audio playback. A "Classify" button triggers the classification process, which uses the trained model to analyze the audio. Once the classification is complete, the result is displayed, showing whether the audio is REAL or DEEPFAKE. Additionally, visualizations such as the waveform and Mel spectrogram are shown to provide more insight into the audio's characteristics.

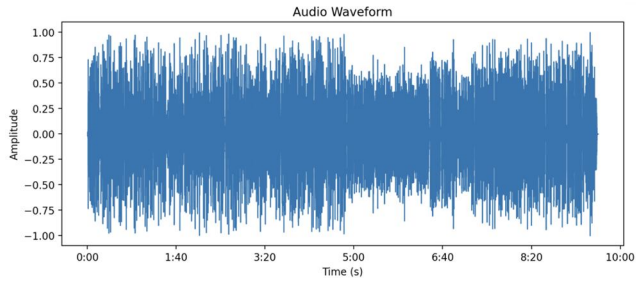
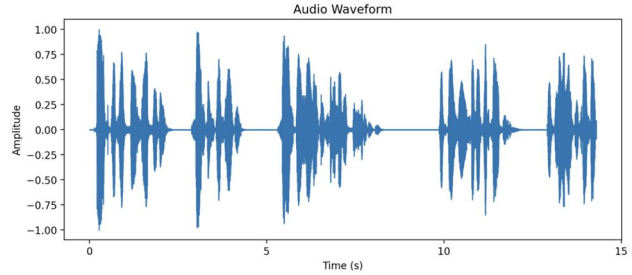
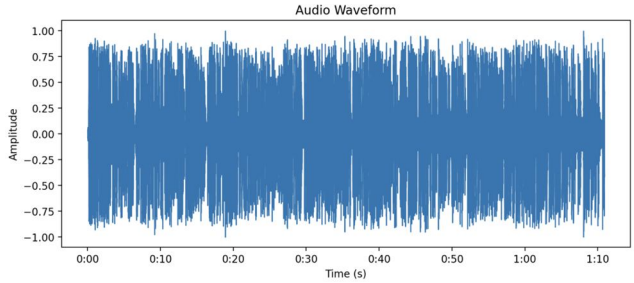
E. Model Storage and Reusability

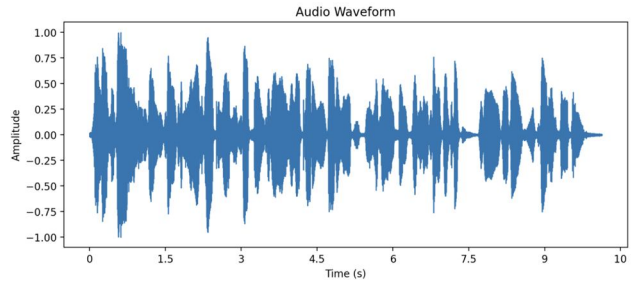
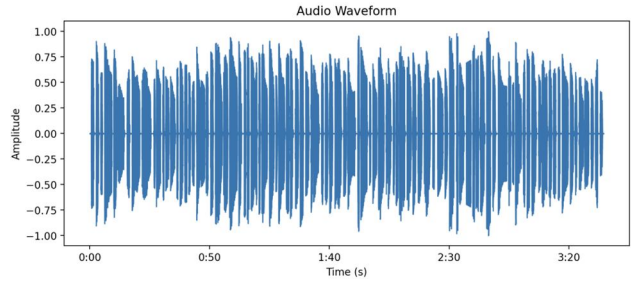
If the pre-trained model is missing, the user can upload it manually through the interface. Once the model is available, it can be loaded and used for predicting whether uploaded audio files are REAL or DEEPFAKE. This ensures that the model can be reused for future predictions without needing to be retrained.

VI. RESULTS AND ANALYSIS

The evaluation of the audio classification system involves analyzing its performance on various test cases. This process compares the system's predictions against the expected outcomes to assess its accuracy and effectiveness. The results of this evaluation are presented in Table 1.

Table 1: Test Cases

Input Audio	Expected Outcome	Predicted Outcome
 sample-1.wav	REAL	REAL
 sample-2.wav	REAL	REAL
 sample-3.wav	DEEPFAKE	DEEPFAKE

 sample-4.wav	DEEPFAKE	REAL
 sample-5.wav	DEEPFAKE	DEEPFAKE

The model's ability to generalize to unseen data is a key factor in its performance. Analyzing the training and validation accuracy can assess how well the model generalizes as it is exposed to more data. A well-generalizing model will have similar performance on both the training and validation sets, indicating that it has learned meaningful patterns from the data, rather than simply memorizing it. This can be observed in Figure 2.

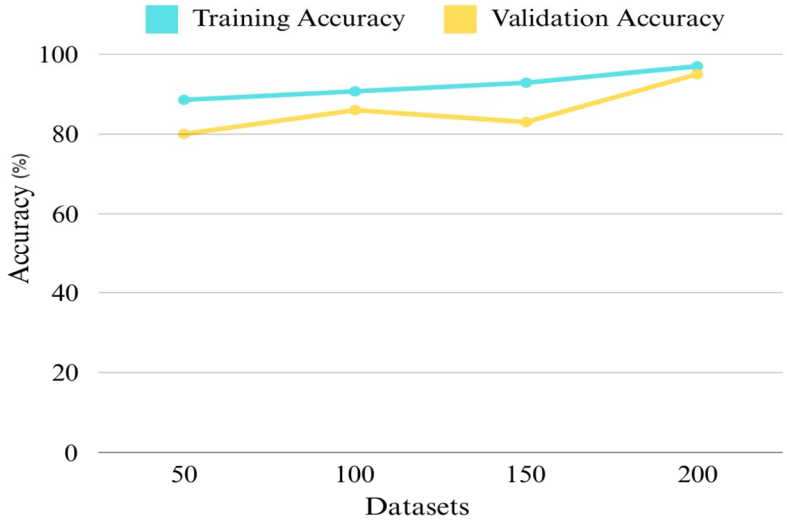


Fig 2: Training and Validation Accuracy

VII. CONCLUSION

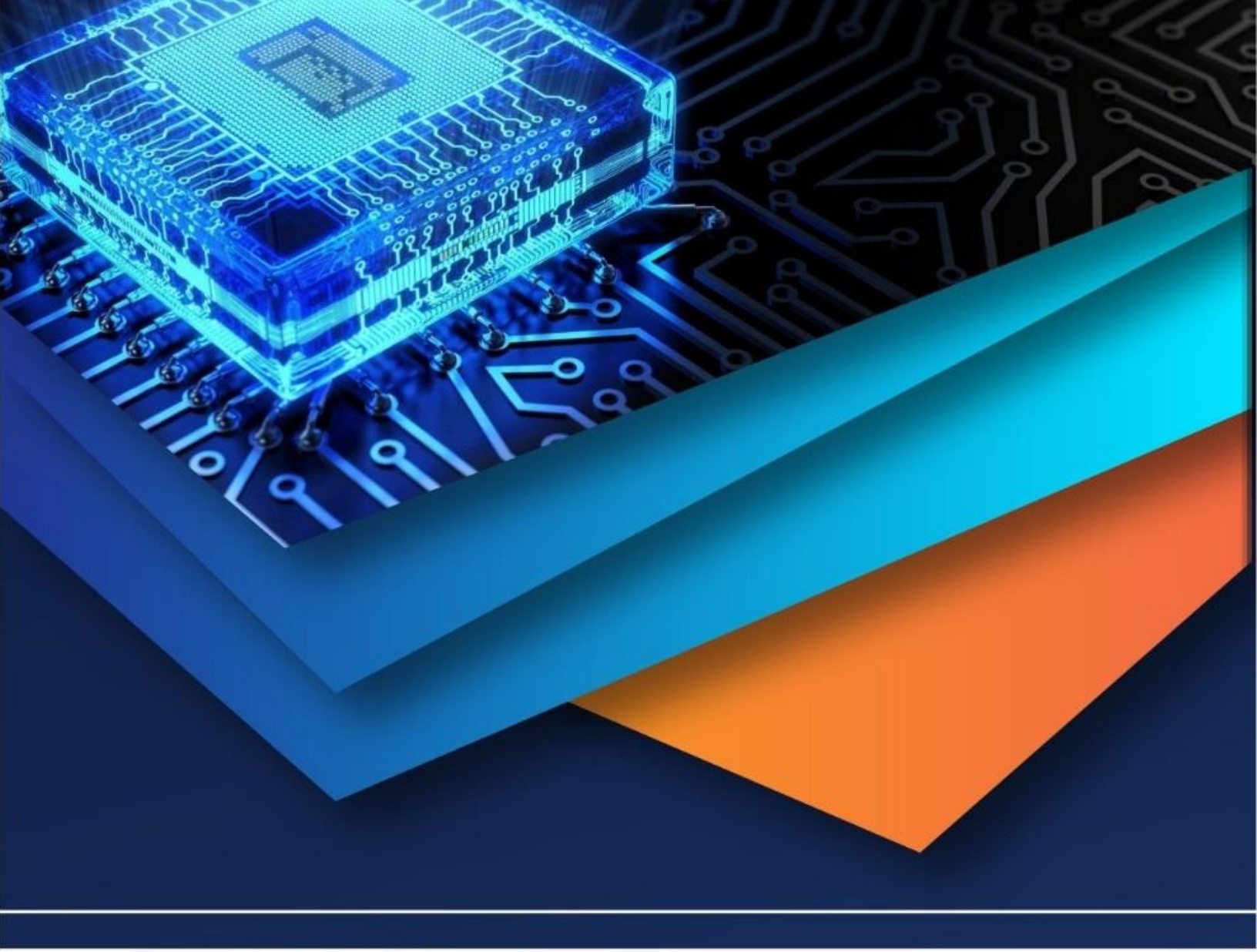
This proposed work provides a foundational framework for detecting deepfake audio by leveraging signal processing and machine learning techniques. By extracting and analyzing meaningful audio features, the system effectively classifies audio as real or synthetic. While the current approach demonstrates robust performance, future work can enhance the framework by incorporating advanced deep learning models, real-time analysis, and support for diverse audio formats. These improvements can broaden the system's applications in areas such as cybersecurity, digital forensics, and media verification, contributing to the ethical use of synthetic audio technologies.

VIII. FUTURE SCOPE

The future scope includes several key improvements. Integrating deep learning models like CNNs or RNNs could enhance detection accuracy by learning hierarchical audio features, surpassing the current SVM classifier. Enabling real-time audio analysis would expand its use in live media verification and fraud detection. Supporting additional audio formats, such as OGG and MP3, would increase versatility across industries. Cross-validation and hyperparameter tuning could improve the model's generalization on unseen data. Advanced feature extraction methods could further enhance accuracy. Additionally, addressing dataset imbalances through oversampling or class weighting would refine the model's performance. These improvements would make the system more robust and applicable to fields like cybersecurity, digital forensics, and media verification.

REFERENCES

[1] Joel Frank and Lea Schönherr, "WaveFake: A Data Set to Facilitate Audio Deepfake Detection," 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks, vol. 6, no. 1, pp. 1-17, 2021.
[2] Suk-Young Lim, Dong-Kyu Chae, and Sang-Chul Lee, "Detecting Deepfake Voice Using Explainable Deep Learning Techniques," Appl. Sci., vol. 12, no. 8, pp. 3926-3940, 2022.
[3] Shaikh Muskan Shaukatali, Parekh Jaini Rajnikant, Chawan Aashrayee Rajan, Bhanushali Aarti Damji and Prof. Salabha Jacob, "Fake Voice Detection System," International Journal for Multidisciplinary Research, vol. 6, no. 2, pp. 1-7, 2024.
[4] Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao, "Audio Deepfake Detection: A Survey," Journal Of Latex Class Files, vol. 14, no. 8, pp. 1-20, 2023.
[5] Brian McFee, Colin Raffel, Dawen Liang, Daniel Ellis, Matt Mcvicar, Eric Battenberg and Oriol Nieto, "Librosa: Audio and Music Signal Analysis in Python," Proc. Of The 14th Python In Science Conf. (SCIPY), pp. 1-7, 2015.
[6] Ameer Hamza, Abdul Rehman Javed, Farkhund Iqbal, Natalia Kryvinska, Ahmad S.Almadhor, Zunera Jalil and Rouba Borghol, "Deepfake Audio Detection via MFCC Features Using Machine Learning" IEEE Access, vol. 10, pp. 134018-134028, 2022.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY