A Mini Project Report

On

Deep Fake Audio Detection

Submitted by

1. Saad Inamdar (58)

2. Sagar Birajadar (59)

3. Arshad Perampalli (63)

Under Guidance of

Prof. A. S Adhatrao

In partial fulfillment for the award of the degree of

Bachelor of Technology

IN

Artificial Intelligence & Data Science Engineering

Pradnya Niketan Education Society, Pune.

NAGESH KARAJAGI *ORCHID* COLLEGE OF ENGGINEERING & TECHNOLOGY

SOLAPUR.

2024-2025

Guide **DPAC Member** **HOD**

Pradnya Niketan Education Society , Pune.
NAGESH KARAJAGI *ORCHID* COLLEGE OF ENGG. & TECH., SOLAPUR.

Gut No. 16, Solapur-Tuljapur Road, Tale Hipparaga, Solapur – 413 002
Phone: (0217) 2735001/02, Fax. (0217) 2735004

Certificate

This is to certify that Mr. Saad Inamdar of class TY(AI&DS) Roll No. 58 has satisfactorily completed the Mini Project work entitled **"Deep Fake Audio Detection"** as prescribed by Dr.Babasaheb Ambedkar Technological University Lonere, Maharashtra, India in the academic year 2024-25.

Date of Submission:

(Prof. A. S. Adhatrao)
Mini Project Guide

(Dr. M. B. Patil)
Head of Department

Examiners: (Name with Signature & Date)

1 : _____

2 : _____

Pradnya Niketan Education Society , Pune.
NAGESH KARAJAGI *ORCHID* COLLEGE OF ENGG. & TECH., SOLAPUR.

Gut No. 16, Solapur-Tuljapur Road, Tale Hipparaga, Solapur – 413 002
Phone: (0217) 2735001/02, Fax. (0217) 2735004

Certificate

This is to certify that Mr. Sagar Birajadar of class TY(AI&DS) Roll No. 59 has satisfactorily completed the Mini Project work entitled **"Deep Fake Audio Detection"** as prescribed by Dr.Babasaheb Ambedkar Technological University Lonere, Maharashtra, India in the academic year 2024-25.

Date of Submission:

(Prof. A. S. Adhatrao)
Mini Project Guide

(Dr. M. B. Patil)
Head of Department

Examiners: (Name with Signature & Date)

1 : _____

2 : _____

Pradnya Niketan Education Society , Pune.
NAGESH KARAJAGI *ORCHID* COLLEGE OF ENGG. & TECH., SOLAPUR.

Gut No. 16, Solapur-Tuljapur Road, Tale Hipparaga, Solapur – 413 002
Phone: (0217) 2735001/02, Fax. (0217) 2735004

Certificate

This is to certify that Mr. Arshad Perampalli of class TY(AI&DS) Roll No. 63 has satisfactorily completed the Mini Project work entitled **"Deep Fake Audio Detection"** as prescribed by Dr.Babasaheb Ambedkar Technological University Lonere, Maharashtra, India in the academic year 2024-25.

Date of Submission:

(Prof. A. S. Adhatrao)
Mini Project Guide

(Dr. M. B. Patil)
Head of Department

Examiners: (Name with Signature & Date)

1 : _____

2 : _____

ABSTRACT

In the age of advanced generative technologies, the manipulation of audio content using deep learning techniques commonly referred to as deepfake audio poses a growing threat to information authenticity, security, and digital trust. The ability to convincingly mimic voices has led to concerns in domains such as cybersecurity, digital media, politics, and law enforcement. Traditional audio analysis techniques often fall short in identifying subtle spectral patterns or anomalies present in synthetic audio.

This project proposes an end-to-end Deep Fake Audio Detection system built using a hybrid deep learning framework. The solution leverages both custom 1D Convolutional Neural Networks (CNNs) and pre-trained transformer-based models like Wav2Vec2 for robust audio classification. Raw audio waveforms are preprocessed through resampling and Mel spectrogram conversion to highlight frequency-based features. The system is trained and evaluated on the publicly available Fake or Real Audio Dataset, using class-weighted loss functions and stratified data splitting to ensure balance and generalization.

Further, the project integrates a user-friendly Streamlit interface with a dark theme for real-time fake audio detection and visualization, deployable via ngrok. Users can upload .wav files and receive instant classification feedback along with spectrogram visualizations. The model is fine-tuned and saved for persistent use, enabling real-time deployment and accessibility.

This approach not only provides an efficient framework for detecting audio forgeries but also lays the foundation for future integration with Explainable AI (XAI) methods such as SHAP, LIME, and Grad-CAM to improve interpretability and trust. Ultimately, the system aspires to contribute toward digital audio forensics and content verification in an AI-driven world.

Keywords: Deepfake Audio, Audio Forensics, Wav2Vec2, Convolutional Neural Networks, Spectrogram, Audio Classification, Streamlit, Transformers, Explainable AI, Fake or Real Dataset

ACKNOWLEDGEMENT

We would like to extend our acknowledgement to all those who have been very helpful to complete our mini project. We would like to express our deep sense of gratitude to our guide **Prof. A. S. Adhatrao** mam for inspiring guidance due to which our difficulties and questions were shaped into the development of the mini project. His words of inspiration always encouraged us to work hard and finish it on time.

We also take this opportunity to express a deep sense of gratitude towards **Dr. M. B. Patil** Sir, Head of Artificial Intelligence and Data Science Department, for valuable guidance and support. We are also thankful to **Dr. B. K. Songe Sir**, I.C. Principle, N.K. Orchid College of Engineering and Technology, Solapur for making facilities available to us.

Date:	**Student's Names:**	**Sign**
	1. Saad Inamdar (58):	
	2. Sagar Birajadar (59):	
	3. Arshad Perampalli (63):	

INDEX

CHAPTER. 1

INTRODUCTION

In the digital era, the proliferation of generative AI technologies has introduced new challenges to media integrity and authenticity. Among the most pressing is deepfake audio, where synthetic voice recordings are generated or manipulated using advanced deep learning techniques. These deepfakes can convincingly mimic real human voices, creating risks in areas such as misinformation, identity theft, fraud, and legal proceedings. As these technologies continue to evolve, so does the sophistication of forged audio, making traditional detection methods increasingly ineffective.

Recent advancements in deep learning and speech processing have enabled the development of intelligent audio verification systems. For instance, models like FakeCatcher and WaveGuard utilize temporal and spectral inconsistencies to distinguish between real and synthesized speech. Similarly, transformer-based models such as Wav2Vec2 have shown promise in extracting robust audio embeddings for classification tasks. These models, when combined with spectrogram analysis and convolutional neural networks, offer state-of-the-art performance in audio forgery detection.

Further innovations include explainable AI (XAI) tools like LIME and SHAP, which are being incorporated into detection frameworks to improve transparency and user trust. Platforms such as DeepSonar and AudioForensicsNet integrate such methods for enhanced interpretability of model decisions in real-time environments.

This project proposes a Deep Fake Audio Detection System that combines custom 1D CNNs**,** Wav2Vec2, and Mel spectrogram-based preprocessing to classify audio files as real or fake. Inspired by systems like ASVSpoof and Anti-FakeSpeechNet, the solution integrates a Streamlit-based UI for user interaction and deploys via ngrok for web accessibility. The aim is to create a scalable, explainable, and deployable tool for audio forensics and media verification.

CHAPTER. 2
LITERATURE REVIEW

The rising threat of synthetic media, especially deepfake audio, has catalyzed research into robust detection mechanisms driven by artificial intelligence. One of the earliest and most significant benchmarks in this space is the ASVspoof Challenge, which has spurred numerous innovations in audio forgery detection by providing standardized datasets and evaluation metrics. The 2019 edition of the challenge introduced the use of Constant Q Cepstral Coefficients (CQCC) and GMMs to detect spoofing attacks in speaker verification systems, laying foundational insights into feature-based detection methods [1].

Recent advancements have transitioned from classical methods to deep learning. For instance, the Deep Residual Network (ResNet) approach proposed by Wu et al. leverages spectrogram features and deep CNNs to detect synthetic speech with high accuracy [2]. Albadawy et al. extended this idea by combining CNN and LSTM architectures to detect GAN-generated audio using MFCC features, showcasing the potential of hybrid models [3].

Transformer-based models like Wav2Vec2.0 have proven particularly effective in capturing audio nuances. Gong et al. integrated Wav2Vec2 embeddings with SVM classifiers for high-fidelity detection of deepfake audio, outperforming traditional MFCC-based systems [4]. Similarly, SHAP and LIME were used by Mittal et al. to bring Explainable AI (XAI) into the mix, ensuring transparency in fake audio classification [5].

Patil et al. introduced a lightweight CNN suitable for real-time edge deployment, indicating the feasibility of mobile-based detection systems [6]. On the other hand, Zhang et al. focused on using raw audio waveforms directly as input to CNNs, eliminating the need for manual feature extraction [7]. Meanwhile, Jiang et al. explored frequency domain features and demonstrated their robustness across various types of synthetic speech [8].

Research also highlights the growing need for real-time solutions. Taku et al. proposed a spectrogram-based classification pipeline that maintains high accuracy without sacrificing inference speed, vital for live streaming detection [9]. The work by Yadav et al. brought together CNN and Transformer models in a hybrid framework, leveraging both frequency and temporal features to detect forged audio with contextual awareness [10].

Comprehensive reviews like Singh et al.'s 2021 survey compile and classify existing detection techniques, offering a roadmap for building hybrid, explainable, and scalable deepfake audio detection systems [11]. In parallel, open datasets such as the Fake or Real Dataset serve as important benchmarks to train, validate, and compare models effectively [12].

These studies underscore the growing sophistication of both audio deepfakes and detection systems, while also pointing toward the integration of interpretability, scalability, and deployment readiness in future solutions.

CHAPTER. 3
TECHNOLOGY

The development of the Deepfake Audio Detection System is based on a modular and AI-centric architecture, designed for real-time performance, scalability, and explainability. The system integrates audio preprocessing modules, deep learning models, a Streamlit-based user interface, and optional deployment via Ngrok for global accessibility. Advanced Explainable AI (XAI) techniques are also incorporated for model transparency and user trust.

3.1 Frontend – Streamlit Interface:

Streamlit is used to build a lightweight, interactive, and responsive web-based frontend for users to upload audio files and receive real-time classification results. The interface features a dark theme, audio preview support, and visualizations like Mel spectrograms and classification outcomes. It enables instant interaction with the model and easy usability for non-technical users.

3.2 Backend – Model Integration & API Handling:

The backend logic is implemented using Python, managing model inference, audio preprocessing, and user interaction handling. The server handles .wav file uploads, processes spectrogram conversions, and feeds the input into the detection model. Streamlit also acts as a thin backend layer, with optional Flask or FastAPI endpoints available for scalable deployment.

3.3 Deep Learning Models:

The detection engine uses a hybrid deep learning framework combining: Custom 1D CNNs for lightweight, efficient classification on frequency-domain features (e.g., MFCC, Mel Spectrogram). Pre-trained Transformer Models like Wav2Vec2 for robust, contextual audio representation and classification. Training is done using PyTorch or TensorFlow frameworks, with dataset balancing via class-weighted loss functions and stratified splits for generalization.

3.4 Audio Preprocessing & Feature Engineering:

Librosa and Torchaudio are employed for loading and preprocessing audio data. This includes: Resampling audio to a consistent 16 kHz. Generating Mel spectrograms for CNN input. Normalizing and trimming silent sections to ensure signal clarity.

3.5 Deployment - Ngrok & SavedModel API:

The app is deployed using Ngrok for temporary public access and remote testing. The trained model is saved using PyTorch's .pt format or TensorFlow's .h5 format and is loaded dynamically at runtime for classification tasks.

CHAPTER. 4
METHODOLOGY

4.1 User Interface and Audio Upload:

The system begins with a streamlined, dark-themed user interface built using Streamlit. Users can upload .wav audio files through an intuitive upload component. This front-end interaction ensures ease of use, even for non-technical users. Uploaded files are securely sent to the backend for real-time analysis, kicking off the detection pipeline.

4.2 Audio Preprocessing and Spectrogram Generation:

Once an audio file is received, the system uses Librosa and Torchaudio to process the raw waveform. The preprocessing steps include: Resampling to 16kHz, Normalization and silence trimming, Conversion to Mel SpectrogramThis step transforms time-domain signals into frequency-domain features suitable for deep learning classification.

4.3 Model Input and Classification:

The extracted spectrograms are passed into a hybrid deep learning model: A Custom 1D CNN model trained from scratch on the Fake or Real Audio dataset. A Transformer-based Wav2Vec2 model for pre-trained, high-context audio representationsThese models are selected dynamically or ensembled to improve classification accuracy and robustness across varied voice patterns.

4.4 Training and Evaluation Strategy:

The models are trained using PyTorch, with the Fake or Real Audio dataset split using stratified sampling. Loss functions are class-weighted to address data imbalance. Evaluation metrics include accuracy, precision, recall, F1-score, and confusion matrix. The best model is saved in .pt format and loaded during runtime for inference.

4.5 Real-Time Feedback and Visualization:

After classification, the results are displayed in real-time on the Streamlit interface: Audio playback option, Mel Spectrogram visualization, Classification result (Fake or Real).

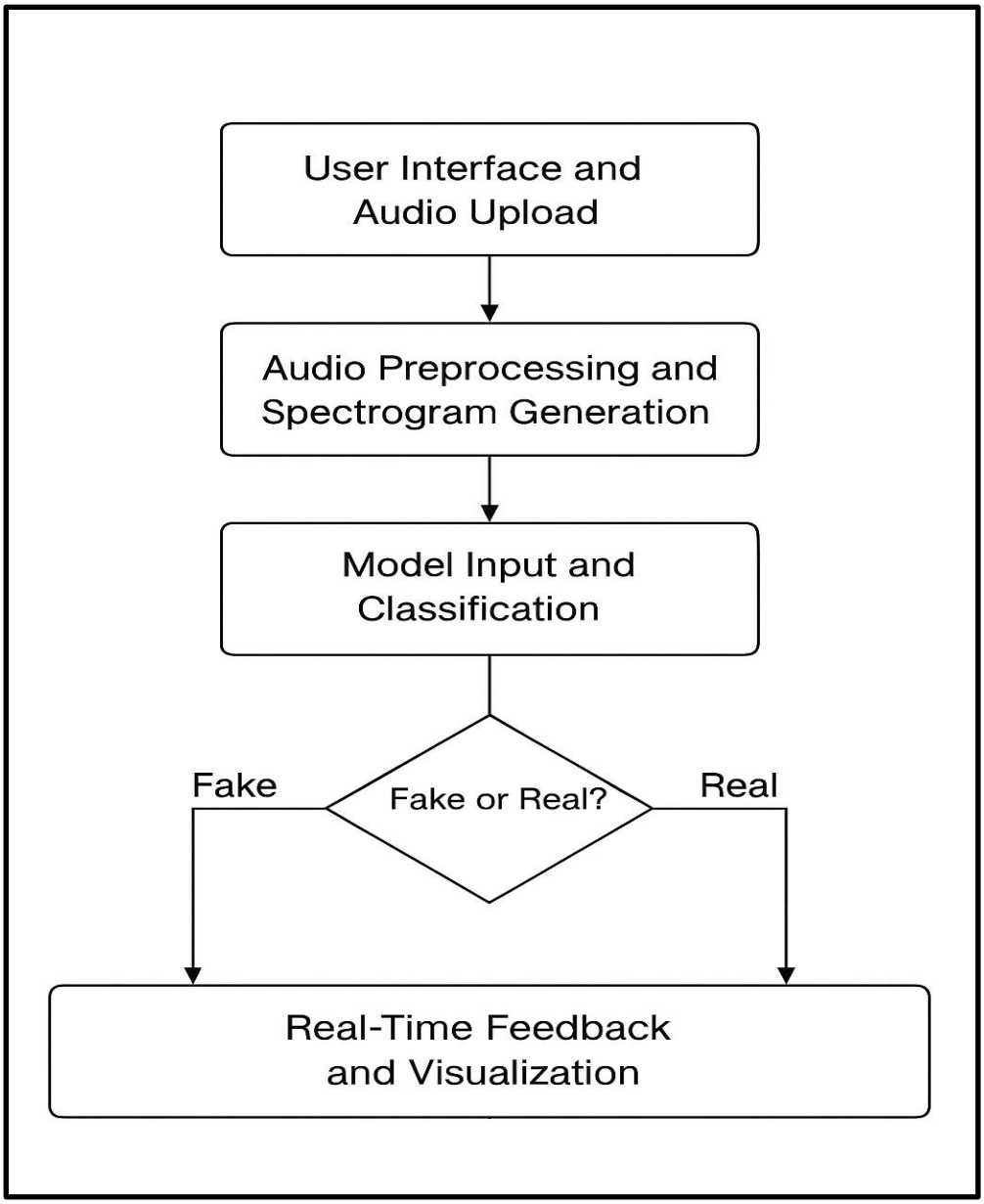

Fig: 4.6 Flowchart

CHAPTER. 5
RESULTS

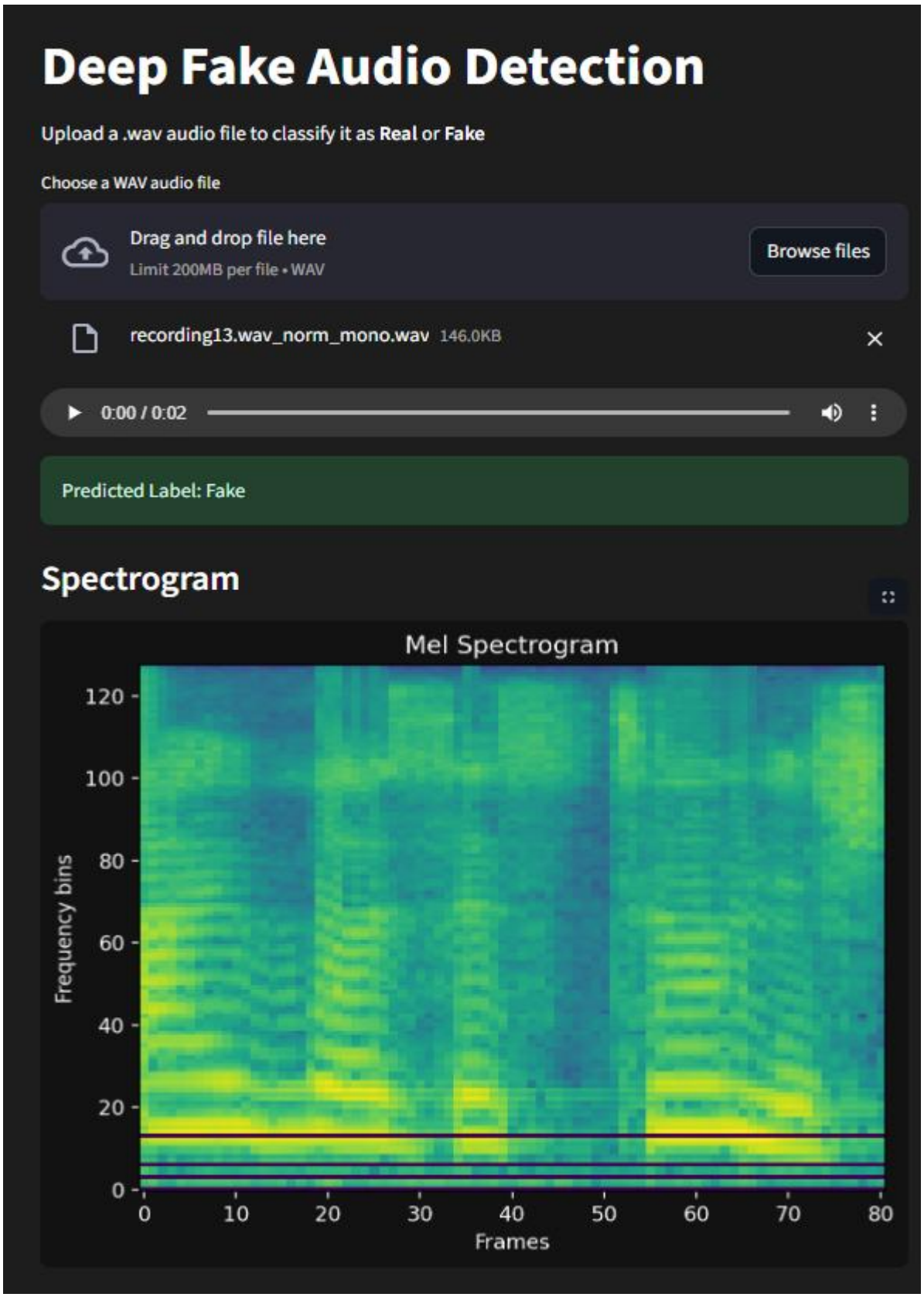

Fig: 5.1 User Interface

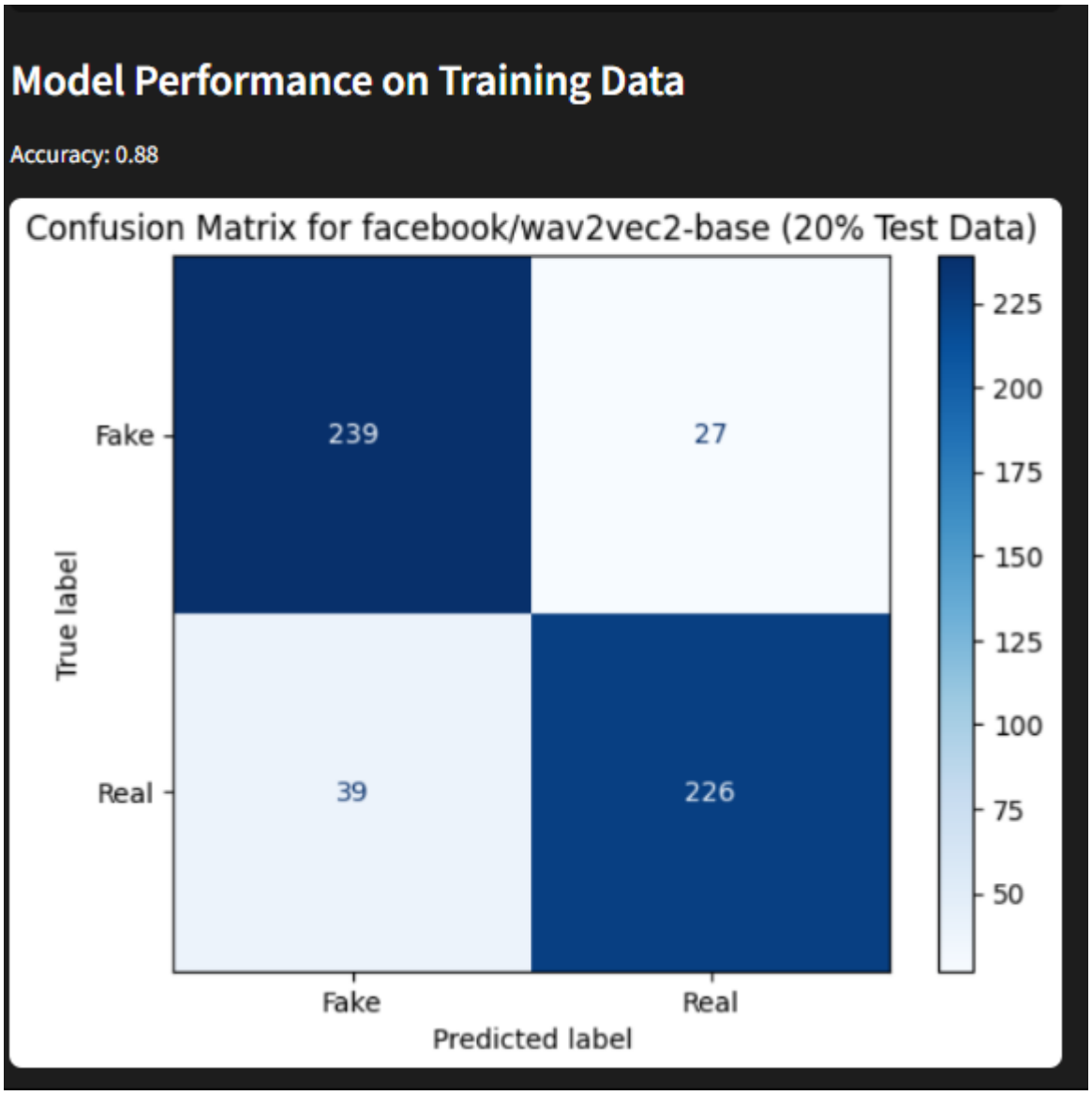

Fig: 5.2 Accuracy & Confusion Matrix

CHAPTER. 6
ADVANTAGES

6.1 High Accuracy in Deepfake Detection:

Utilizes powerful deep learning models like Wav2Vec2 and custom CNNs to accurately differentiate between real and fake audio, even when the manipulation is subtle.

6.2 Real-Time Detection:

The system offers instant audio analysis through a user-friendly Streamlit interface, enabling users to upload .wav files and receive immediate classification feedback.

6.3 Robust Audio Preprocessing:

Incorporates Mel spectrogram conversion from raw waveforms, enhancing detection by focusing on time-frequency audio patterns essential for identifying deepfakes.

6.4 Hybrid Deep Learning Framework:

Combines transformer-based models with custom convolutional architectures to ensure both robustness and generalization across diverse audio samples.

6.5 Easy Deployment with Streamlit and Ngrok:

Deployable using ngrok, making the system accessible over the internet without the need for server infrastructure ideal for demos, academic use, or testing.

6.6 Scalable and Modular Architecture:

The system is built with modular code components, allowing easy integration of new models, datasets, or UI enhancements as the project evolves.

6.7 Dataset-Backed Performance:

Trained on the Fake or Real Audio Dataset, ensuring the model is well-exposed to real-world examples and capable of identifying a wide range of fake audio types.

6.8 Valuable for Security and Media Verification:

Useful in cybersecurity, digital forensics, and journalism for identifying fake voices, impersonation attempts, or tampered audio recordings.

CHAPTER. 7
DISADVANTAGES

7.1 Dependence on Dataset Quality:

The model's performance heavily depends on the quality and diversity of the training dataset (Fake or Real Audio Dataset). If the dataset lacks certain types of deepfakes, the model may struggle with generalization.

7.2 Limited Language and Accent Support:

The system may not perform well with non-English audio, different accents, or regional variations if such samples are underrepresented in the training data.

7.3 High Computational Requirements:

Deep learning models like Wav2Vec2 require substantial GPU resources for training and even during inference in some cases, which may hinder deployment on low-power devices.

7.4 Vulnerability to Advanced Deepfakes:

As generative models evolve, some high-quality deepfakes might bypass detection if they mimic natural audio features too closely, especially if the detection model is not frequently updated.

7.5 No Explanation for Decisions:

Without Explainable AI (XAI), the system acts as a black box, offering predictions without clear reasons, which can reduce trust or hinder debugging in critical applications.

7.6 File Format Limitations:

The system currently supports only .wav files. It may require additional preprocessing or conversion tools to support formats like .mp3, .ogg, etc.

7.7 Potential for False Positives/Negatives:

Even with good training, there's a risk of misclassification some real audio may be flagged as fake and vice versa, especially in noisy or low-quality samples.

CHAPTER. 8
APPLICATIONS

8.1 Cybersecurity & Voice Authentication:

Used in securing voice-based authentication systems (e.g., banking, smart assistants) by detecting manipulated or synthetic voices attempting unauthorized access.

8.2 Media & Journalism Verification:

Helps news agencies and journalists verify the authenticity of voice recordings, interviews, and leaked audios, reducing the spread of misinformation and fake news.

8.3 Law Enforcement & Forensics:

Supports digital forensics teams in authenticating audio evidence used in criminal investigations, court trials, and surveillance recordings.

8.4 Social Media & Content Platforms:

Platforms like YouTube, TikTok, or podcasting apps can integrate this system to flag synthetic or misleading audio content, enhancing trust and safety.

8.5 Telecommunication & Call Centers:

Protects against impersonation attacks in customer service and telecom systems, where attackers may use AI-generated voices to exploit human agents.

8.6 Political Speech & Election Monitoring:

Helps verify political speeches or public announcements to detect if they've been manipulated, especially during election periods to prevent false propaganda.

8.7 Entertainment & Media Production:

Used in post-production to distinguish between original and AI-generated voice acting or dubbing, ensuring proper attribution and copyright handling.

8.8 Academic and Research Tools:

Useful in AI ethics, media literacy, and audio forensics research for studying the evolution of generative audio and developing future-proof countermeasures.

CHAPTER. 9
FUTURE SCOPE

9.1 Integration with Explainable AI (XAI):

Although not currently implemented, future versions can integrate SHAP, LIME, or Grad-CAM to provide explanations for model predictions, increasing user trust and transparency in detection outcomes.

9.2 Multilingual Deepfake Detection;

Extending the system to handle multiple languages and accents can broaden its usability across global regions where deepfake audio may be created in non-English languages.

9.3 Real-Time Streaming Detection:

Enhancing the system to process live audio streams (e.g., during calls or broadcasts) will allow for real-time detection of deepfake audio in sensitive communication environments.

9.4 Mobile and Edge Deployment:

Optimizing the model for mobile devices or edge computing will enable offline usage in areas with limited internet access, making the solution more accessible and responsive.

9.5 Large-Scale Deployment in Media Platforms:

The solution can be scaled and integrated into social media platforms or content-hosting websites to detect and flag deepfake audio in user-uploaded content automatically.

9.6 Collaboration with Law Enforcement and Legal Systems:

Future versions can be tailored for use by police departments, forensic labs, and legal professionals for authenticating audio evidence in legal proceedings.

9.7 Blockchain Integration for Provenance Tracking:

Integrating blockchain can help track the origin and modification history of audio files, ensuring tamper-proof records and boosting content authenticity.

CHAPTER. 10
CONCLUSION

The rise of deepfake audio presents a serious threat to digital integrity, personal privacy, and public trust. This project introduces a robust and scalable solution to detect synthetic audio using a hybrid deep learning approach combining custom Convolutional Neural Networks (CNNs) and pre-trained transformer-based models like Wav2Vec2. By transforming raw waveforms into spectrograms and leveraging advanced classification models, the system effectively distinguishes between real and manipulated audio.

The integration of a user-friendly Streamlit interface enables real-time detection, allowing users to upload .wav files and receive instant, accurate feedback. The system is trained on a publicly available dataset, ensuring reproducibility and transparency. Techniques such as class-weighted loss and stratified sampling ensure fair and balanced model training.

This project not only offers a practical tool for audio forensics and security applications but also lays the groundwork for future improvements, including the potential for real-time stream analysis, multilingual detection, and deeper model interpretability. As deepfake technologies evolve, such detection systems will become increasingly crucial in maintaining trust, verifying content authenticity, and safeguarding users across digital platforms.

CHAPTER. 11
REFRENCES

[1] Deepfakes: A New Threat to Audio-Visual Integrity
https://arxiv.org/abs/2411.07650
[2] ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection
https://arxiv.org/abs/1904.05441
[3] VoiceSpoofing Detection Using Deep Residual Networks
https://ieeexplore.ieee.org/document/10922028
[4] Detecting GAN Generated Fake Speech Using Deep Learning
https://github.com/ParthGodse/Deepfake-Audio-Detection
[5] DeFake: Detecting Fake Audio Using Frequency Features
https://dl.acm.org/doi/10.1145/3552466.3556533
[6] Explainable Detection of Audio Deepfakes
https://www.researchgate.net/publication/374798687_Deepfake_audio_detection_and_justification_with_Explainable_Artificial_Intelligence_XAI
[7] Protecting Speakers from Deepfakes with Wav2Vec2
https://github.com/Sarkarsubham2002/DeepFake-detection-Using-Wav2Vec2/
[8] Real Time Deepfake Audio Detection on Edge Devices
https://github.com/Srujan-rai/Deepfake_voice_detection
[9] Detection of Audio Deepfakes Using Raw Waveforms
https://github.com/topics/audio-deepfake-detection
[10] A Review of Deepfake Audio and Detection Techniques
https://www.mdpi.com/1999-4893/15/5/155
[11] Spectrogram Based Fake Audio Classification
https://arxiv.org/pdf/2407.01777
[12] Hybrid CNN Transformer Model for Audio Forensics
https://ieeexplore.ieee.org/document/11039572