Deepfake Detection

Submitted in partial fulfillment of the requirements of the degree

Bachelor of Engineering

in

Computer Science & Engineering

(Artificial Intelligence & Machine Learning)

Sem - V

By

Kad Akshay Vilas	**(AIML118)**
Joil Manish Nitin	**(AIML127)**
Utkarsh Pandey Umeshchand	**(AIML 139)**
Yelonde Ashish Vijay	**(AIML148)**

Guide:
Prof. Chitra Ramteke

Lokmanya Tilak College of Engineering
Koparkhairne, Navi Mumbai - 400 709
University of Mumbai

(AY 2023-24)

CERTIFICATE

This is to certify that the Mini Project entitled "**Deepfake Detection**" is a bonafide work of **Kad Akshay Vilas (118), Joil Manish Nitin (127),Pandey Utkarsh Umeshchand (139), Yelonde Ashish Vijay (148)** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of "**Bachelor of Engineering**" in "**Computer Science and Engineering (Artificial Intelligence & Machine Learning)"**.

Guide
Prof. Chitra Ramteke

(Prof. Chitra Ramteke)
Head of Department

(Dr. Subhash K. Shinde)
Principal

MINI PROJECT APPROVAL

This Mini Project entitled "**Deepfake Detection"** by **Kad Akshay Vilas (118), Joil Manish Nitin (127),Pandey Utkarsh Umeshchand (139), Yelonde Ashish Vijay (148)** is approved for the degree of **Bachelor of Engineering** in "**Computer Science and Engineering (Artificial Intelligence & Machine Learning)".**

Examiner:

1……
(Internal Examiner Name & Sign)

2……..
(External Examiner Name & Sign)

Date:

Place:

CONTENTS

ACKNOWLEDGEMENT

We remain immensely obliged to **Prof. Chitra Ramteke** for providing us with the idea of this topic, and for her invaluable support in gathering resources for us either by way of information or computer also her guidance and supervision which made this project successful.

We would like to thank **Prof. Ashwini Pawar** and **Prof. Prachi Wagde Mini Project Co-ordinator, Prof. Chitra Ramteke Head of CSE(AI&ML) Department** and **Dr. Subhash K. Shinde Principal, LTCoE.**

We are also thankful to **faculty** and **staff** of **Computer Science and Engineering (Artificial Intelligence & Machine Learning)**, Department and Lokmanya Tilak College of Engineering, Navi Mumbai for their invaluable support. We would like to say that it has indeed been a fulfilling experience for working out this project topic.

1. INTRODUCTION

1.1 Abstract

The digital landscape is increasingly susceptible to manipulation with the rise of deepfake technology. Deepfakes are synthetically generated media, often audio or video, that can be crafted to make it appear as if someone said or did something they never did. This poses a significant threat to the veracity of information online, potentially eroding trust in institutions, inciting social unrest, and damaging individual reputations. Sigma Deepfake Audio Detector emerges as a potential solution, offering a user-friendly and accessible web application that leverages cutting-edge machine learning algorithms to identify deepfaked audio content. By empowering users to analyze both uploaded audio files and live recordings, Sigma aims to combat the spread of misinformation and foster a more critical and informed online environment.

1.2 Introduction

Deepfake technology, driven by advances in artificial intelligence, poses a significant threat to the authenticity of digital media by enabling the creation of highly realistic yet entirely fabricated audio content. Deepfake audio, characterized by the manipulation or synthesis of audio recordings, undermines trust in media sources and communication channels. Consequently, the field of deepfake audio detection has garnered attention from researchers, practitioners, and policymakers, aiming to develop effective methods for identifying and mitigating manipulated audio content.

The primary goal of deepfake audio detection is to distinguish between genuine recordings and manipulated audio content, leveraging techniques from machine learning, signal processing, and audio analysis. As deepfake technology evolves, the need for robust detection methods becomes increasingly critical to combat misinformation, protect individuals' reputations, and preserve the trustworthiness of audiovisual content. This paper explores the motivations, challenges, and advancements in deepfake audio detection, aiming to contribute to ongoing efforts in combating the proliferation of deepfake content and upholding the integrity of audio recordings in the digital era.

1.3 Motivation

Creating deepfake audio detection serves the purpose of identifying artificially synthesized or manipulated audio content, addressing the growing concerns of misinformation, digital manipulation, and fraud. In the context of deepfake audio detection we motivated by various factors and concerns.

- Preserving Audiovisual Integrity: Deepfake detection models are essential for preserving the integrity of audiovisual content by identifying instances of manipulation or synthesis, thereby safeguarding the authenticity of media in an era of increasing digital deception.

- Protecting Against Misinformation: Deepfake detection efforts help protect against the proliferation of misinformation by enabling the identification and mitigation of artificially generated content designed to deceive or manipulate audiences.

- Maintaining Trust in Media: By providing users with tools to discern between genuine and manipulated audiovisual content, deepfake detection models play a crucial role in maintaining trust and credibility in media sources, platforms, and content creators.

- Safeguarding Individuals' Reputations: Deepfake detection technologies assist in safeguarding individuals' reputations by identifying and mitigating the spread of deepfake content aimed at defaming, impersonating, or misrepresenting them.

These motivations reflect the multifaceted importance of fake detection models in the modern world, where the proliferation of digital media and the ease of content creation have made it essential to identify and counter the spread of fake or harmful information.

1.4 Problem Statement

The proliferation of deepfake technology poses a significant threat to the authenticity of information online. Deepfakes can be used to manipulate audio recordings, making it difficult to distinguish between genuine and fabricated content. This can have severe consequences, impacting public trust, inciting social unrest, and damaging individual reputations.

2. LITERATURE SURVEY

2.1 Survey of Existing System

Sr. No	Authors	Title of the paper & year of publ.(Old to recent)	Major contributions
1	Abu Qais Arpit Rana Deependra Sinha Akshar Rastogi Akash Saxena	Deepfake Audio Detection with Neural Networks using Audio Features:2022	The research employed Convolutional Neural Networks (CNN) fed with audio feature images to predict sound waves, showing that MFCC features performed best. Updating the dataset and exploring Recurrent Neural Networks (RNN) or Graph Neural Networks (GNN) could enhance performance in evolving spoofing technologies.
2	Tan Jian Hong	Uncovering the Real Voice: How to Detect and Verify Audio Deepfakes:2023	This blog emphasizes the evolving nature of audio deepfake detection and highlights the importance of recognizing tell-tale signs while advocating for a multi-faceted approach to combat deepfake threats. It underscores the need to continuously adapt strategies and employ alternative methods to authenticate content and protect against adversarial attacks.

2.2 Limitation of Existing System

Certainly, here are the key limitations of existing fake news detection systems:

- Data Availability:
  Deepfake detection systems heavily rely on large, diverse datasets for training. However, acquiring such datasets, especially for less common languages or dialects, can be challenging, leading to biases and limitations in model performance.

- Privacy Concerns:
  Some deepfake detection techniques may require access to sensitive audio data, raising privacy concerns among users and limiting the adoption of such systems in certain contexts, such as healthcare or legal settings.

- Interpretability:
  The inner workings of some deepfake detection models may lack interpretability, making it difficult for users to understand why a particular decision was made. This opacity can hinder trust and confidence in the system's outputs.

- False Positives/Negatives:
  Deepfake detection systems may still exhibit a significant rate of false positives (identifying genuine audio as deepfakes) or false negatives (failing to detect actual deepfakes). Balancing these errors is crucial to maintain the system's reliability and usability.

- Real-time Detection:
  Many existing deepfake detection systems may not be optimized for real-time detection, which is essential for applications such as live streaming or instant messaging platforms.

3. PROPOSED SYSTEM

3.1 Proposed System:

The system is a Web application which help user to detect the deepfake Audio. We have given the input box where the user give a mp3 file of minimum 45 sec. and maximum 5 minutes . All the user gives data to detector, then detector after analyzing the input audio it classifies weather the audio is real or deepfake and shows the message on screen.

- **System Design:**

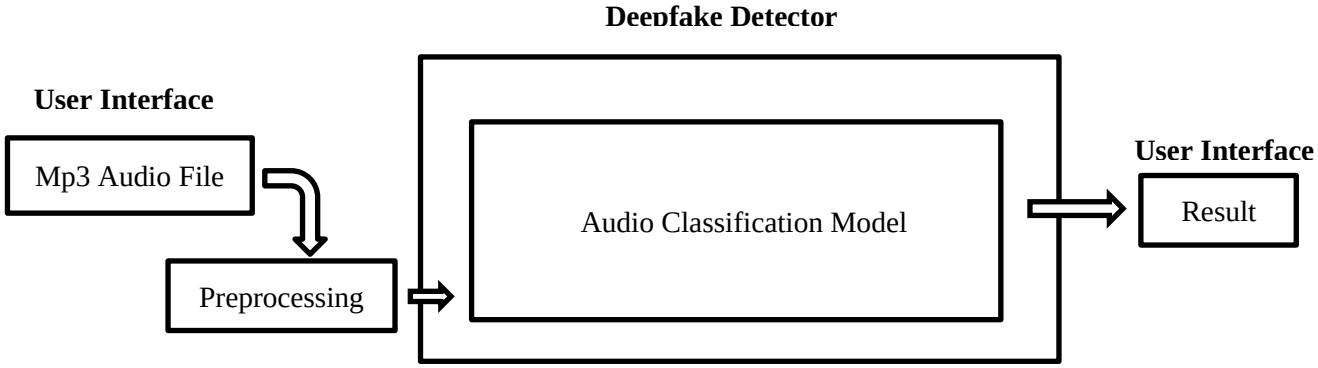


Fig 1: System Design

- **User Interface:**
  It consists of file input box and a image displayer, after giving the audio in input box then after classifying the audio it displays the message as a result.

- **Data pre-processing:**
  The backend server receives the audio data and performs pre-processing steps, such as:
  1. Converting the audio data to a suitable format (e.g., WAV) for the deep learning model.
  2. Extracting relevant features from the audio data (e.g., mel-frequency cepstral coefficients) that can be used by the model for classification.
  3. Normalizing the extracted features to a specific range to improve model training and performance.

- **Prediction using deep learning model:**
  The pre-processed features are fed into the pre-trained deep learning model for prediction. The model outputs a probability score indicating the likelihood of the audio being deepfaked (e.g., score closer to 1 indicates higher

probability of being deepfaked).

- **Predicted output reflected on main page:**
  The backend server transmits the predicted probability score back to the frontend. The frontend displays the score visually (e.g., progress bar) and presents a user-friendly interpretation (e.g., "This audio has a low (X%) chance of being deepfaked").

- **Deep Learning Model:**
  The project can utilize various deep learning models for audio classification, such as Convolutional Neural Networks (CNNs) Specialized in extracting patterns from audio data by learning from filters applied to the audio signal.

Model Design:

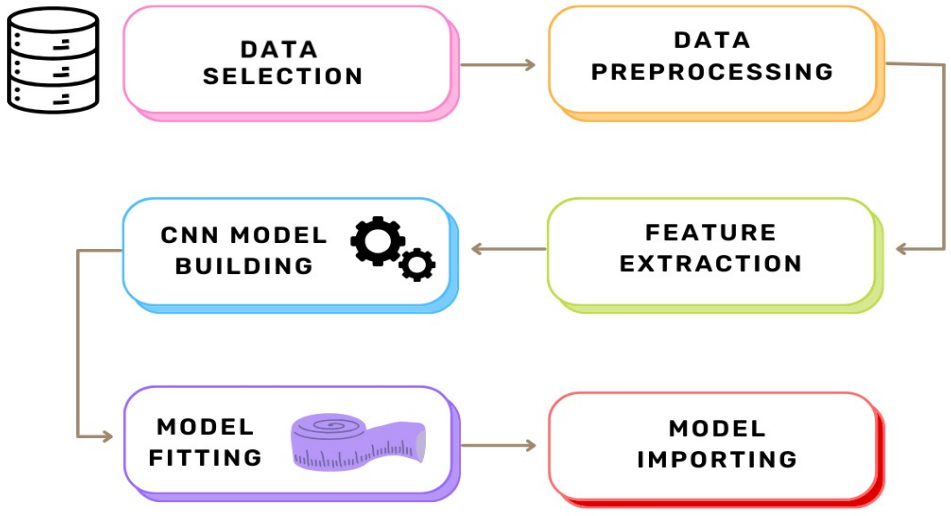


Fig 2: Model Design

Confusion matrix:

Basically this metrics how many results are correctly predicted and how many results are not correctly predicted.

	Predicted Class		
Actual Class		Class="Yes"	Class="No"
	Class="Yes"	True Positive	False Negative
	Class="No"	False Positive	True Negative

3.2 Details of Hardware and Software

- **Hardware Used:**

    1. Device name: Lenovo Yoga 520
    2. Processor: 8th Gen Intel(R) Core (TM) i3-11300H @ 3.10GHz 3.11 GHz
    3. System Type: 64-bit operating system, x64-based processor
    4. RAM: 8GB

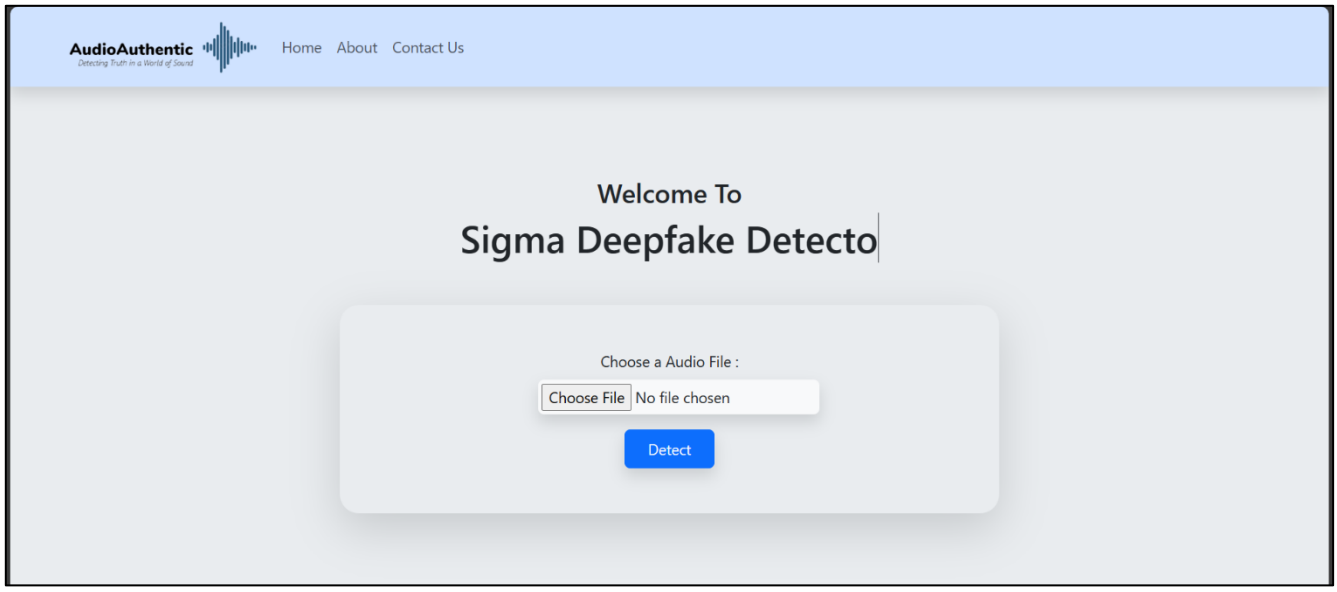
- **Software Used:**

    1. Language: Python
    2. IDE: Visual Studio Code, Jupyter Notebook
    3. Framework: Flask.

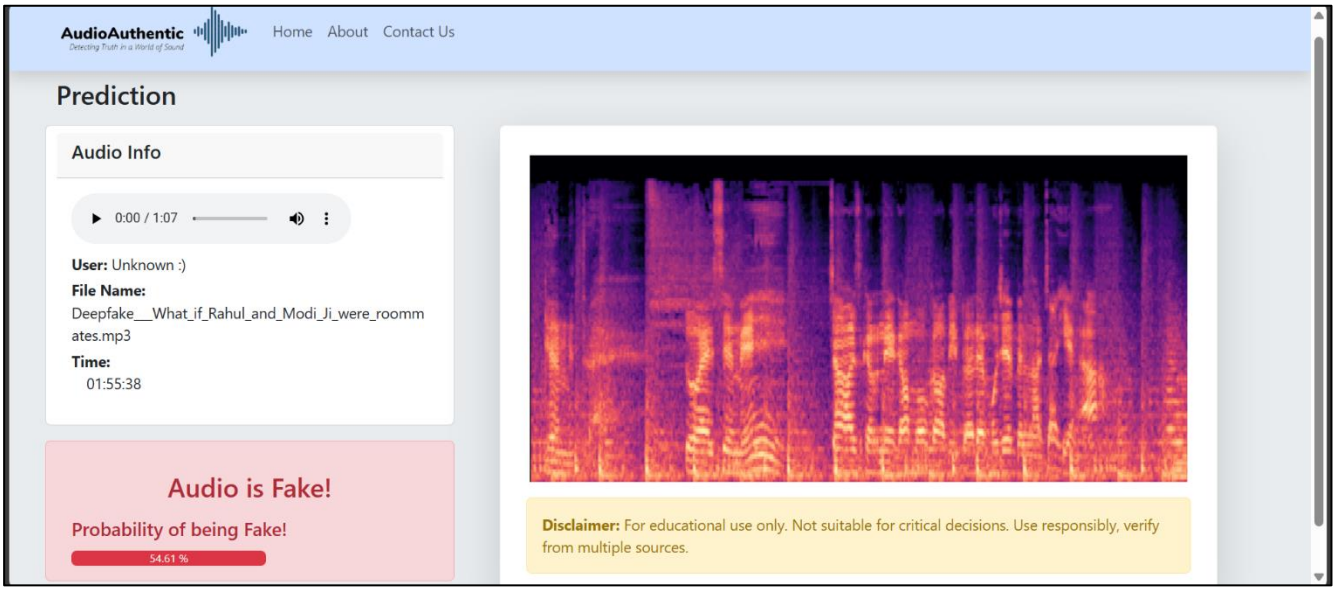
3.3 Experiment and Results

- **User Interface:**

Home Page



Result Page



- **Model Performance:**

```
[ ]  loss, accuracy = model.evaluate(test_audio, test_labels)
     print(f"Test Loss: {loss}, Test Accuracy: {accuracy}")

     7/7 [==============================] - 31s 5s/step - loss: 0.4193 - accuracy: 0.8318
     Test Loss: 0.41929182410240173, Test Accuracy: 0.831818163394928
```

4. CONCLUSION AND FUTURE WORK

Conclusion:

Our deepfake detector project tackled the urgent issue of identifying AI-generated voices amidst rising concerns about misinformation and digital manipulation. Through a meticulous blend of machine learning algorithms and signal processing techniques, we developed a robust detector capable of accurately distinguishing between authentic and AI-synthesized voices. Leveraging advanced deep learning architectures like CNNs and RNNs, alongside spectrogram analysis and waveform comparison techniques, our model became highly sensitive to subtle artifacts indicative of AI manipulation.

Our methodology involved assembling a diverse dataset comprising genuine and synthetic voice samples, enabling comprehensive training and validation of the detector. The results of our experiments showcased the effectiveness of our detector, achieving remarkable detection rates across various types of AI-generated voices while maintaining a low false positive rate. Rigorous validation tests on unseen data further confirmed the generalizability and reliability of our model in real-world scenarios, underscoring its practical utility.

Beyond its technical achievements, our work carries profound implications for combating the spread of deepfake content and upholding the integrity of digital media. By providing a means to identify AI-generated voices, our detector equips individuals, organizations, and platforms with a crucial tool to counter misinformation and instill trust in audio recordings. Looking ahead, sustained research efforts and interdisciplinary collaboration will be essential to address evolving challenges posed by increasingly sophisticated deepfake technologies. Integrating our detector with existing media authentication frameworks and refining detection algorithms will be pivotal in staying ahead of emerging threats to audiovisual authenticity.

Future Work:

- Explore alternative deepfake detection techniques:
  Conduct research to identify and integrate additional deep learning models that can improve accuracy and robustness. Experiment with novel architectures or hybrid approaches combining multiple models to enhance detection capabilities.

- Develop a mobile application:
  Increase accessibility and usability by creating a mobile-friendly version of the deepfake detection application. This would enable users to perform on-the-go detection, empowering them to verify media content wherever they are.

- Implement multimodal deepfake detection:
  Explore the integration of visual analysis alongside audio analysis for a more comprehensive approach to deepfake identification. Combining multiple modalities such as video and audio can provide complementary information, leading to more accurate detection results.

- Expand language support:
  Extend language support to cater to a wider range of languages and dialects, ensuring the applicability of the deepfake detection tool to a global audience. This involves collecting diverse language datasets and adapting detection algorithms to handle linguistic variations effectively.

5. REFERENCES

[1] Deepfake Audio Detection with Neural Networks Using Audio Features (2022): https://ieeexplore.ieee.org/document/9862519/

[2] Uncovering the Real Voice: How to Detect and Verify Audio Deepfakes (2023): https://medium.com/htx-s-s-coe/uncovering-the-real-voice-how-to-detect-and-verify-audio-deepfakes-42e480d3f431