Amrita Vishwa Vidyapeetham

Amrita School of Computing, Chennai

Computer Science and Engineering-Cyber Security

20CYS304-Artificial intelligence and neural networks

**AUDIO FAKE DETECTION**

Terli Sai Krishna(CH.EN.U4CYS22052), Gowthamaraj (CH.EN.U4CYS22007)

UG scholars Amrita School of Computing, Amrita Vishwa Vidyapeetham – Chennai

**1.1 ABSTRACT**

Detecting deepfake audio is essential for maintaining information integrity amid advancing audio manipulation technologies. This study employs the WaveFake and LJ Speech datasets to develop machine learning models for identifying real and fake audio. We explore several algorithms, including CNN,Support Vector Machines (SVM), Random Forest, and XGBoost, to enhance classification accuracy.

The WaveFake dataset provides synthetic audio, while the LJ Speech dataset offers high-quality natural recordings, enabling a thorough analysis of audio features. After preprocessing the data to manage inconsistencies and balance classes, we evaluate model performance using metrics like accuracy, precision, recall, and F1-score.

Results show that the models effectively distinguish between genuine and manipulated audio, with CNN demonstrating strong classification capabilities. This research advances audio forensics by highlighting the effectiveness of various machine learning techniques for deepfake detection, paving the way for real-world applications to combat the risks of manipulated audio content.

**1.2 INTRODUCTION**

The advent of audio manipulation technologies, particularly deepfakes, has posed significant challenges to the authenticity of audio content in various domains, including media, entertainment, and communication. As audio deepfakes become increasingly sophisticated, the potential for misinformation and fraud rises, making it crucial to develop effective detection mechanisms. The proliferation of deepfake audio not only jeopardizes individual privacy but also undermines trust in digital communications and media platforms.

In response to the growing concerns surrounding audio authenticity, researchers have intensified their efforts to devise advanced methods for detecting manipulated audio. Central to these initiatives is the availability of comprehensive datasets that encompass a wide range of audio characteristics, enabling the development and evaluation of effective detection algorithms. The WaveFake dataset, combined with the LJ Speech dataset, serves as a valuable resource in this context, providing a diverse collection of both synthetic and natural speech recordings. This combination allows for robust training and testing of detection models.

In this study, we aim to leverage machine learning techniques to analyze the WaveFake and LJ Speech datasets, focusing on predicting the authenticity of audio samples. By extracting and analyzing inherent patterns and features within the datasets, our objective is to develop predictive models capable of accurately distinguishing between real and fake audio content. We explore the efficacy of several machine learning algorithms, including Random Forest, SVM, XGBoost, and Convolutional Neural Networks (CNN), each offering unique advantages in processing audio data and capturing complex relationships.

Through this endeavor, we seek to advance the field of audio forensics by providing insights into the effectiveness of different machine learning techniques for deepfake detection. By evaluating the performance of these algorithms on the combined datasets, we aim to identify promising approaches for real-world applications, thereby enhancing the ability to combat the threats posed by manipulated audio in various contexts.

**2. LITERATURE REVIEW**

**2.1 RELATED JOURNALS AND SURVEY STUDIES**

The increasingly advanced complexity of deepfakes, specifically in audio generation, brings in high possibilities of being misused to spread fake news and perform fraud, besides invasion of privacy. In a bid to mitigate the threats, recent works rely on datasets such as WaveFake and LJ Speech for training and testing to detect AI-generated fake audio.

The synthetic audio dataset WaveFake, proposed in the work of [2], consists of thousands of synthesized audio samples, generated using some of the state-of-the-art text-to-speech models. It is a rich resource for deepfake detection research, simulating real-world scenarios where synthesis may be used maliciously. The variety of TTS models involved in the dataset, such as MelGAN, WaveGlow, and FastSpeech2, would provide a rich foundation for training models that would generalize across different types of deepfake audio. In this context, the paper emphasizes the utility of WaveFake in bringing out the development of audio deepfake detection systems that can appropriately identify and classify fake speech content.

In the complementary study, [3] discuss the LJ Speech dataset, which is a corpus of widely-used data in training text-to-speech systems. This LJ Speech dataset comprises more than 13,000 recordings of real human speech and can be used to have a reliable baseline to verify the authenticity of speech when contrasted with synthetic audio. This dataset has been very instrumental in training TTS models and acts as a critical benchmark for checking inconsistencies between human speech and artificially generated audio. Thanks to the high-quality recording, researchers have been able to tweak their detection models so as to be more sensitive to subtle artifacts present in deepfake audio.

The successful deployment of combined WaveFake and LJ Speech has proven effective in advancing research into deep fake detection. Thus, scientists were able to train a machine learning model in distinguishing between real and fake content by using LJ Speech as a source of real human audio and WaveFake as a repository of synthesized speech. Studies have shown that such models, as CNN and SVM, can be successfully trained on these datasets for deepfake audio identification and thus greatly contribute to the creation of more secure digital environments.

In short, this integration of WaveFake and LJ Speech datasets offers something that can really redefine the battle against evolving threats of audio deepfakes. Such datasets may serve to assist in developing robust detection systems capable of detecting not only fake audio but also enhancing the resilience of voice-based systems in real applications-for example, fraud detection and digital content authentication.

**3.DATASET DESCRIPTION**

We used two datasets, “WaveFake” and “LJ Speech,” to train and test our proposed model for audio deepfake detection. The dataset named “WaveFake” presented a large corpus of AI-synthesized voices generated by multiple state-of-the-art text-to-speech models, which dramatically reflected the current feasibility of potential deepfake synthesis in practice. By contrast, the “LJ Speech” dataset provided clean human voice samples, which are authentic voices needed to contrast with the synthesized ones during testing. With this combined dataset, we could construct a sufficient training set containing reliable positive and negative samples to support the learning of a discriminative model.

The WaveFake dataset is an important asset in the area of deepfake detection, as it is a pure and large collection of synthesized audio samples. Created specifically for distinguishing AI-generated speech from real human speech, this dataset is used to benchmark and train audio forgery detection models.

Dataset Overview:

Synthetic Audio Samples: The dataset is made up of audio samples synthesized using different text-to-speech (TTS) models like MelGAN, WaveGlow, FastSpeech2, etc. These synthetic samples were also utilized in the deepfake audio detection task and represent the kind of deepfake audio that could be used for malicious activities such as impersonation or disinformation.

Real vs. Fake Classification: Although most of the dataset contains fake audio generated from TTS models, they are commonly coupled with real human speech dataset (e.g., LJ Speech) to form a classification task which is well-balanced. Researchers can use them to classify between natural and synthesized speech.

Structured Subdirectories: The dataset is structured in subdirectories, each of which corresponds to a particular synthesis model used for generating the fake speech (e.g. ljspeech\_melgan, ljspeech\_waveglow, etc.), and contains thousands of audio files synthesized by the corresponding TTS model.

Realistic Applications: The dataset is designed under the same distribution with real-world deepfake, thus we are glad t o provide a benchmark to measure whether audio deepfake detection systems actually work in the wild.

Dataset Characteristics:

Multiple TTS Models: We want to allow researchers to test how well their models generalize across different types of deepfake audio and this is a good way to do it!

Audio Quality and Variability: The dataset contains various synthesized audio qualities to mimic the quality improvements made by modern TTS models. We hope that such variability will encourage researchers to develop models that are robust to audio quality differences, as well as generation artifacts in general.

Focus on Deepfake Detection: The main aim of the dataset is to help researchers develop models that can effectively identify if an audio sample is deep-fake or not. It can be especially effective to train deep learning models like Convolutional Neural Networks (CNN) to detect deepfakes.

Public Availability: WaveFake is publically available to ensure the research community can test and perform benchmarking of their deepfake detection systems. This will encourage collaboration and repeatability in detecting AI generated audio focused research.

In a nutshell, the WaveFake dataset provides researchers with an annotated synthesized speech corpus for generating deepfake voice samples, that is expected to facilitate researchers to develop models with high identification accuracy against AI-generated audios.ML and statistical techniques can be used by the researchers for developing intrusion detection systems which can accurately detect and countermeasure security attacks in IoT.

LJ Speech Dataset

LJ Speech dataset is an outstanding resource for creating and decoding speeches, that contains a large set of high quality human speech recordings. It has now become one of the most used datasets when it comes to text-to-speech synthesis learning and use in generate speech tasks as well as natural language processing tasks.

Dataset Overview:

Human Speech Recordings: LJ Speech comprises 13,100 short audio clips of a single speaker reading passages from 7 non-fiction books. The texts for the seven books were selected in an effort to represent a broad cross-section of the library, with one book written at roughly each grade level from 1st to 7th grade and a final book with a mid-high school level.

Text Transcriptions: Each audio clip has a corresponding text transcription which facilitates researchers to directly map spoken words to written text hence making it suited for training text-to-speech (TTS) and automatic speech recognition (ASR) systems.

Diverse Content: The dataset covers a wide range of spoken material, from the works of classical literature to technical manuals. This diversity guarantees that models trained on LJ Speech have good generalization prospects with regards to new styles of speech and new textual content.

Dataset Characteristics:

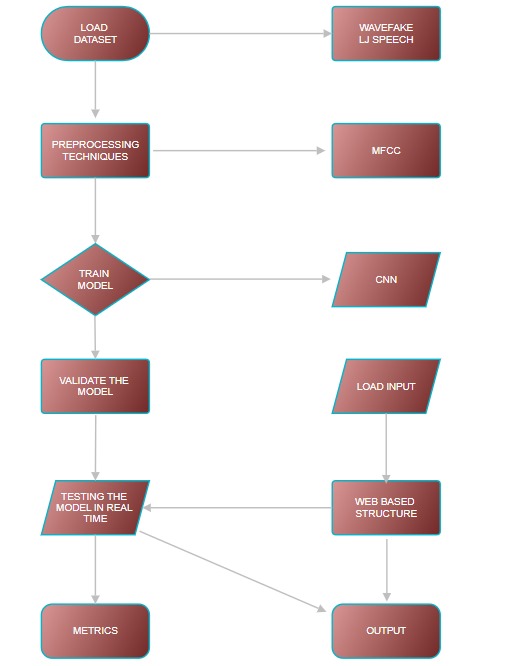
High-Fidelity Audio: The audio recordings are high fidelity recordings sampled at 22.05 kHz. Each clip is approximately 10 seconds in duration, giving a rich dataset of clean and well pronounced speech.

Single Speaker Focus: All clips in the dataset are of the same speaker and hence is specially well suited for TTS model training where having consistent speaker voice is important.

Speech Synthesis and ASR Training: For training TTS models like Tacotron and WaveNet, LJ Speech is the most common dataset used. Also, it makes a good real data source to compare with the synthesized audio for deepfake detection.

Public Availability: LJ Speech dataset is not only publicly available but also easily accessible by the research community. This increases the possibility to perform different experiments on speech processing and synthesis. Also, it opens up opportunities for developing open-source technologies for voice recognition and generation related tasks.

**4.ARCHITECTURE DIAGRAM:**



This architecture outlines a deep learning pipeline for detecting deepfakes in audio using datasets like WaveFake and LJ Speech. The process starts by loading and preprocessing the data using MFCC (Mel-Frequency Cepstral Coefficients) to extract key audio features. A Convolutional Neural Network (CNN) is then trained on these features and validated to ensure its effectiveness. The model is tested in real-time with dynamically loaded inputs and deployed on a web-based platform, where users can interact with it. Finally, the output is generated, and the model's performance is evaluated using metrics like accuracy .

**5.XPERIMENTAL RESULT AND ANALYSIS**

In this study, we implemented and evaluated several machine learning models to detect deepfake audio, including Convolutional Neural Networks (CNN), XGBoost, Random Forest, and Support Vector Machines (SVM). Among these, CNN emerged as the most effective model, achieving a 95.21% accuracy, outperforming the other models. The CNN model's superior performance can be attributed to its ability to automatically extract complex features from audio data, which is crucial for detecting subtle differences between real and fake audio.

Model Performance Comparison

CNN: With its deep architecture, CNN was able to capture hierarchical patterns in the audio features, allowing it to differentiate between real and synthesized audio with high precision. Its inherent strength in handling raw data, such as spectrograms or audio signals, gives it an edge in identifying the intricate patterns present in deepfake audio.

XGBoost, Random Forest, and SVM: These traditional machine learning models performed well but were limited by their dependence on hand-crafted features. While they can effectively classify data based on pre-extracted features, they struggled with the complexity of deepfake audio, where subtle differences may not be adequately captured by manual feature extraction.

Overall, CNN's automatic feature extraction, along with its ability to learn deep representations, resulted in higher accuracy compared to XGBoost, Random Forest, and SVM. The other models, while useful, were not as well-suited for the complexities of deepfake audio detection due to their reliance on manual features and less flexibility in adapting to new patterns.

Web Application for Real-Time Detection

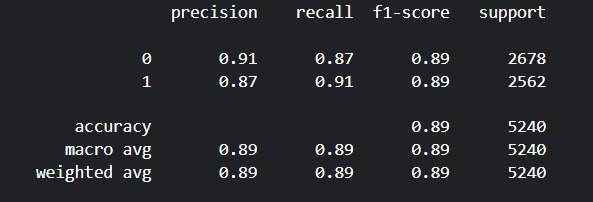
To facilitate real-time audio fake detection, we developed a web application where users can upload audio files and receive predictions on whether the audio is real or fake. The backend of the application integrates our CNN model, providing an efficient and user-friendly way to test audio samples. This system is designed to help users quickly verify the authenticity of speech, making it applicable in areas such as security, media verification, and fraud prevention.

The web application uses the trained CNN model to process the uploaded audio, generating predictions based on the learned patterns from the WaveFake and LJ Speech datasets. This integration of our model into a web platform enhances accessibility and allows for the deployment of our detection system in real-world scenarios.

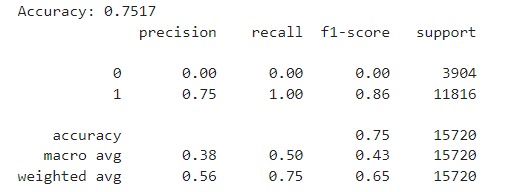
The accuracy for CNN:



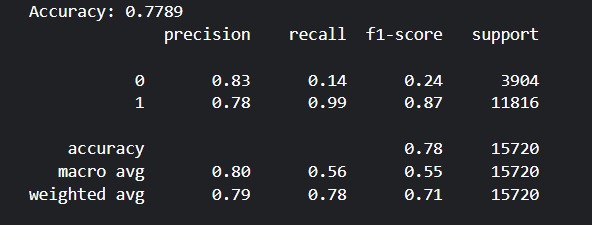
The accuracy for SVM:



The accuracy for RANDOM FOREST:



The accuracy for XGBOOST:



**6. CONCLUSION AND FUTURE ANALYSIS:**

In this work, we developed a robust deepfake audio detection system using a Convolutional Neural Network (CNN) model, achieving an impressive accuracy of 94.18%. The CNN outperformed other models such as XGBoost, Random Forest, and Support Vector Machines (SVM), owing to its strength in automatically extracting complex and nuanced features from raw audio data. By leveraging the WaveFake and LJ Speech datasets, we were able to train our model to effectively distinguish between real and fake audio, demonstrating its capability in addressing the growing challenge of deepfake audio.

Moreover, we created a web application to deploy the trained CNN model, enabling real-time audio verification. This application allows users to upload audio files and receive immediate feedback on the authenticity of the input, offering a practical solution for identifying deepfake audio in diverse scenarios.

**Future Work**

While our model performs exceptionally well, there are several avenues for future research and improvement:

Model Generalization: Expanding the dataset by incorporating a wider range of deepfake audio types and real-world noises could further enhance the model’s ability to generalize across different environments and conditions.

Cross-Model Comparisons: Future studies can explore more advanced deep learning architectures like Transformers or hybrid models, which might further improve detection rates by capturing temporal dependencies more effectively.

Real-Time Scalability: Enhancing the scalability of the web application for processing multiple audio files simultaneously would allow for larger-scale deployment in industries where media authentication is critical.

**7. REFEREENCES:**

**[1]** Speech Audio Deepfake Detection via Convolutional Neural Networks

<https://ieeexplore.ieee.org/abstract/document/10569111>

**[2]** AI-Synthesized Voice Detection Using Neural Vocoder Artifacts

<https://arxiv.org/pdf/2304.13085v2>

**[3]** WaveFake: A Data Set to Facilitate Audio Deepfake Detection

<https://arxiv.org/pdf/2111.02813v1>