National College of Ireland

Forensic Analysis of Synthetic Speech: A Hybrid Approach to Audio Deepfake Detection

MSc Research Project
MSc Cybersecurity

Avani Naidu
Student ID: 23285044

School of Computing
National College of Ireland

Supervisor: Raza Ul Mustafa

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Avani Chandrashekhar Naidu
Student ID:	23285044
Programme:	MSc in Cybersecurity **Year:** 2024-2025
Module:	Practicum
Supervisor:	Raza Ul Mustafa
Submission Due Date:	11-08-2025
Project Title:	Forensic Analysis of Synthetic Speech: A Hybrid Approach to Audio Deepfake Detection
Word Count:	7956 **Page Count** 20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Avani Chandrashekhar Naidu

Date: 11-08-2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Table of Contents

Forensic Analysis of Synthetic Speech: A Hybrid Approach to Audio Deepfake Detection

Avani Naidu
23285044

Abstract

The spread of artificial speech created by highly sophisticated AI methods posed considerable issues in cybersecurity and forensics. In this work, a hybrid forensic model to detect and analyze deepfake audio is introduced, taking into account systematic differences in acoustic and visual aspects in a speech recording. With the dataset ASVspoof2019 Logical Access, we generate Mel-Frequency Cepstral Coefficients (MFCCs) and mel-spectrogram to create a Random Forest (RF) classifier and a Convolutional Neural Network (CNN), respectively. The RF model gives an interpretable answer in terms of SHAP value analysis, whereas the CNN is used through Grad-CAM to find localizable discriminative spectral characteristics pointing to the presence of spoofed speech. In order to make detection more robust, we add more measures of voice quality beyond PESQ, e.g., jitter, shimmer, pitch variability, and harmonics-to-noise ratio that are routinely used in speech pathology, but are not widespread in spoofing detection pipelines. Our proposed hybrid strategy shows an improved rate of accuracy and interpretability when it comes to differentiating authentic and spoofed sound. The presented framework runs through a Streamlit browser-based interface, which allows real- time audio signals, model decision visualization, and clear forensic investigation. This study is of value in the emerging discipline of explainable AI in digital forensics, that paired with a reasonable and explainable solution is effective in mitigating audio-based cyber threats.

1 Introduction

1.1 Background

The growth of artificial intelligence and generating modeling in particular has brought a variety of types of synthetic media with deep spoof audio being of the most concern. The creation of such deepfake audio clips is provided with skills like Text-to-Speech (TTS), Voice Conversion (VC) and high-end generative models like GANs and WaveNet. Although at the moment deepfake voice was created with both accessibility and entertainment in mind, currently, it presents a severe danger in cybersecurity as well, as voice authentication hacking, social engineering, and misinformation attacks are the mentioned areas that stemmed out of deepfake audio **[1][2].**

Current models, like WaveNet, Tacotron 2 and VITS have made it possible to create speech that is almost identical to those of humans, even to those who are intimately familiar with the voices of humans in the real world **[**3**][**4**]**. The shortcomings of current detection systems have been brought to the fore in this conversion of synthetic speech into a cyber-defense hazard, particularly criminal and legal applications where visibility and means of tracing evidence are paramount **[5].**

To meet this challenge, this project proposes a hybrid forensic detection framework that seeks to detect and interpret synthesized audio in the form of a combination of spectrogram-based deep learning, and structured acoustic features. Using such datasets as ASVspoof2019

Logical Access **[6]**, this paper aims at investigating the classification of bona fide and spoofed speech based on Convolutional Neural Networks (CNN) applied to mel-spectrograms and the Random Forest (RF) classifiers applied to MFCCs. Moreover, explainable AI(XAI) tools like SHAP and Grad-CAM are also included to strengthen the forensic value of the models so that decisions made using them are transparent **[7][8]**.

1.2 **Motivation**

The rationale behind this study is the increasing unauthorized use of artificial speech in real world cyber mediated threats. As deepfake audio technology advances, and as audios of actual people become easier to access, voice recognition systems and forensic practices are becoming easily dupable **[6]**. The problem is not only the ability to successfully detect spoofed speech, but also present an understandable and evidence-based argument to each of the detections, which is essential when computers are used in an investigatory or in a legal capacity.

Current machine learning models tend to be a black box where little understanding can be derived as to why a specific audio sample was categorized as spoof. This interpretability makes them weak as far as the forensic applications are concerned. Moreover, numerous systems yield no generalization to previously unobserved spoofing attacks or even have no deployment stages **[9].**

By presenting a clear, understandable, and implementable detection scheme, this project will resolve the issues that arise along with it. A combination of CNNs and RF models makes it possible to conduct a dual-view analysis, including high-dimensional Spectro-temporal patterns and interpretable acoustic properties. The application of SHAP to feature importance and Grad-CAM to visualize spectrogram guarantees that corresponding model decision is traceable, verifiable, and interpretable thus making the system more applicable in a forensic and cybersecurity context **[7][8]**.

1.3 **Research question and Objectives**

"**How can forensic approaches in cybersecurity be applied to detect and interpret deepfake audio used in cyber-enabled threats?**"

After having an extensive review of the existing papers on the topic i.e., synthetic speech cyber enabled threats and how It can be handled with digital forensic approach in cybersecurity. My goal through this project is to explore the methods of how forensic method can be applied to detect and interpret deepfake audio. My research objectives are as follows:

- To investigate existing literature on deepfake audio generation, its misuse in cybercrime, and forensic approaches for its detection and analysis.
- To propose a forensic-oriented framework that enables the detection and interpretation of synthetic speech in cybersecurity contexts.
- To design and evaluate models trained on structured representations of audio for identifying deepfake speech.
- To apply model interpretability mechanisms to analyze the reasoning behind detection decisions in a forensic context.
- To assess the effectiveness of the proposed framework in supporting cybersecurity investigations involving spoofed audio.
- To draw conclusions based on the findings and provide recommendations for future work in synthetic speech forensics.

1.4 Scope and Limitations

1.4.1 Scope

The proposed project is dedicated to the forensic detection of deepfake (synthetic) audio via a hybrid method based on the utilization of both the classical machine learning approach and the deep learning one. In particular, it is designed to deal with logical access attacks, in which synthesized speech is created on Text-to-Speech (TTS) and Voice conversion (VC) solutions. The system can sort out audio samples of bona fide or spoofed using:

- Convolutional Neural Networks (CNNs) based on Mel-spectrograms of high-dimensional pattern recognition.
- Handcrafted features, including MFCCs, pitch-related statistics, were used as an input to Random Forest classifiers.
- Explainable AI (XAI) techniques such as SHAP and Grad-CAM to get insightful and interpretable explanations why predictions were made.

The model is trained and evaluated on the ASVspoof2019 Logical Access as one of the commonly-used benchmark datasets in the study of deepfake deception by voice. The project will also incorporate the design of a prototype interface (where applicable) through Streamlit in order to spoof the real-life deployment. In general, this research will offer scalable, interpretable, and forensically valid solution to the issue of the presence of synthetic speech.

1.4.2 Limitations

Although it provides useful receipts into the practical and academic fields, the project has a number of drawbacks:

- **Constraints in Dataset:** Being trained on ASVspoof2019 LA dataset, the system is tested and validated with it, yet, the dataset does not represent all the possible spoofing modalities, leaving newer ones (such as VALL-E or cross-lingual synthesizers) unrepresented. This has the potential to have an effect on the generalization of unseen deepfake methods.
- **Logical Access-Focused:** The implementation of the design is only logical access-focused (pre-recorded or synthesized audio attacks). It does not manage physical access attacks like replay attacks or live impersonation scenarios and would need more preprocessing and countermeasures.
- **Real-time Capability:** Current solution can not be optimized to run in real-time or low-latency since a prototype interface is created instead. The nature of processing spectrograms and SHAP explanations will cause latency, so they may not be used in applications requiring immediate action, such as biometric authentication.
- **Explainability Scope:** Both SHAP and Grad-CAM are successful in interpretation but have narrow scopes in interpretation since they are restricted to feature-level or visual activation explanations. They are wholly ineffective at capturing language or semantic clues that might be used to distinguish deepfakes in conversational scenario.
- **Audio Quality Sensitivity:** The audio might damage the model-based on poor audio quality and noise and especially where the similarities between the audio in the training data do not include environmental distortions.

1.5 Structure of the Report

The report starts by introductory remarks that give a description of the backdrop and drive to carry out a study in deepfake audio forensics. It describes information about the current problems of deepfake audio in cybersecurity, the research problem and objectives, the scope, and limitations of the study. This is preceded by a literature review, which discusses the

literature behind audio deepfake generation, detection techniques and the use of explainable AI (XAI) in the forensic realm. In this section, the weakness of existing methods is outlined and the necessity of such a hybrid and explainable detection framework is argued.
In the research methodology section, I have given a description of the mixed method used in the research work. It specifies the dataset (ASVspoof2019 Logical Access), characteristics of feature extraction (mel-spectrograms and MFCCs), a classification model (CNN, Random Forest), and explainability tools (SHAP and Grad-CAM) integration. The experimental environment and procedure of training of the model have also been described in the section. The implementation section explains the practical requirements of the project setup. It is composed of the environment setup, tools, Python libraries, and the development process. Reproducibility is supported by screenshots, code snippets and visual outputs.
Performance metrics of models in terms of accuracy, precision, recall and F1-score and confusion matrix analysis are provided in the results and evaluation section. It also comprises visualizations of SHAP values and activation maps of Grad-CAM that explain the model decision on particular test cases.
The results are elucidated and the forensic significance of the results is discussed in the discussion section. It also looks at the advantages and constraints of the proposed framework and gives an insight on how it works to respond to the challenges revealed in literature review.
Lastly, the report provides a conclusion by stating the major findings and contributions of the study. It also contains pointers toward future research directions including enhancing generalizability across settings of spoofing and combining physical access attacks as well as real-time detection.

2 Related Work

2.1 Overview

In recent years new methods of producing highly realistic-sounding synthetic speech, colloquially known as deepfake audio, using artificial intelligence, have exploded onto the scene. Such techniques are based on Text-to-Speech (TTS) and Voice Conversion (VC) models, which rely on deep neural networks and models of generation such as WaveNet, Tacotron 2, and VITS. Initial research was largely about improving the naturality of synthetic voices so that they could be used to access and be entertained. Yet, malign application of this technology has put the emphasis on the detection and forensic procedure.
Some of the researchers have attempted solutions to identify spoof audio signal processing techniques, statistical model, and machine learning approaches. The ASVspoof challenge series **(2015, 2019, and 2021)** also helped tremendously in field development since they offered standardized data and benchmarks to spoof detection models. **[9]** Presented CQCC-GMM as a baseline system, but subsequent research has focused more deep learning-based classifiers, especially Convolutional Neural Networks (CNNs) that find and extract information in mel-spectrograms, which perform better.
Audio-based features have been joined by other metadata features, i.e., item duration, energy distribution, and pitch variation, in order to enhance classification performance. Previous studies by **[10]** have focused on handcrafted features and classic classifiers such as Random Forests especially in low-resource environments, where deep models are simply too compute-intensive.

2.2 Existing Solutions and Technologies

There is now a variety of technical solutions to detect synthetic speech. These can be divided into two broad categories: the deep learning-based methods and the traditional machine learning based methods which are based on feature.

Common models used in deep learning audio training tasks include CNNs and RNNs on audio signals as mel-spectrograms. These graphics mostly capture changes that can be related to time and spectral changes which more normally range varied in bona fide to spoofed speech. CNN-based architectures, use a variety of convolution layers in order to discover the spoofing patterns which cannot be detected by human hearing. Grad-CAM support allows visualizing the model activation to understand what section of the spectrogram contributed to the conclusion by a forensic analyst.

Conversely, more popular traditional machine learning models, such as Support Vector Machines (SVM) or Random Forests (RF), use such handcrafted features as MFCCs, pitch, and energy assignations. The approaches are computationally less demanding and they could be better on a smaller dataset they can generalize. Random Forest has good baseline performance and it is both extensible and highly interpretable, especially when combined with explainability tools such as SHAP.

Further, explainable AI (XAI) has been beginning to rise in popularity in synthetic speech detection. Such tools as SHAP (SHapley Additive Explanations) include model-agnostic explanations which aids in the interpretation of black-box model predictions. In the field of forensics and legal applications, transparency and interpretability is essential and these tools are particularly valuable.

Although there are sophisticated models and tools in the offing, deployment ready solutions are few and far between. The vast majority of models are learned and evaluated on restricted domains and have little flexibility to variations in real-world data sets like noise, compression, or an accent.

2.2.1 Evolution of Deepfake audio and Deepfake Detection

The technology of deep learning has jumped forward, which has shown its marks in the field of speech synthesis and voice conversion technologies. Recent generative approaches, like WaveNet **[1]** and Tacotron 2 **[11]**, or unified in ones like VITS **[12]**, have enabled the generation of synthetic speech that sound almost the same as natural human speech, both prosodically and clearly. These technologies were initially designed to support assistive uses and accessibility applications but due to the current technological advances, they have been largely abused negatively in malicious areas like fraud, impersonation, and disinformation. The development of these audio deepfakes has created the need of strong cybersecurity and forensic science countermeasures.

The ASVspoof Challenge has been a cornerstone in advancing spoof detection research. The ASVspoof2019 Logical Access (LA) dataset, in particular, includes a broad array of synthesized and converted attacks designed to test countermeasure robustness under diverse conditions **[9].**

2.2.2 Acoustic Feature-Based Detection Approaches

Initial anti-spoofing systems mostly used manually designed acoustic attributes including Mel-Frequency Cepstral Coefficients (MFCC), Constant-Q Cepstral Coefficients (CQCC), and Linear-Frequency Cepstral Coefficients (LFCC). These were normally used together with Gaussian Mixture Models (GMMs) to carry out classification. As an example, Todisco et al. showed that CQCC-GMM models represented a moderate performance on already known attacks, with little generalization to unknown methods of spoofing. **[9].**

2.2.3 **Deep Learning with Spectrograms and CNN Architectures**

Yielding to this trend of deep learning, researchers switched to convolutional networks (CNN) with mel-spectrograms or log-mel spectrograms as input features. Lavrentyeva et al. presented a Light CNN (LCNN) architecture, which out-performed the classical models on ASVspoof2019-LA. Subsequently, more time and frequency resolution were implemented by using ResNet, MobileNet, and attention layers. Although these improvements are good, the problem is they are black boxes and this cannot be accepted in forensic and legal practice where explainability is essential. **[13]**.

2.2.4 **CNNs on Mel-Spectrogram**

The better accuracy of Convolutional Neural Networks (CNNs) in the detection of deepfake audio has been revealed on mel-spectrogram representations in recent studies. **[14]** instantiated a model-based CNN training on images of mel-spectrograms, to recognize Bona fide and spoofed speech. Their model successfully extracted local and global features in Spectro-temporal domain- e.g. artifacts of text-to-speech (TTS) and voice conversion (VC) attacks, reflective of the good pattern recognition performance of CNNs in the domain of audio forensics **[14].** Equally, **[15]** published a CNN baseline that used log-mel spectrograms, which achieved better results, gaining an advantage over conventional spectral representations, such as MFCC and CQCC. With the features of a two-dimensional discrete cosine transform (2D-DCT), their detection errors were greatly reduced on the ASVspoof 2019 benchmark **[15]**.

The results confirm the use of CNNs in this project and, in particular, their features to automatically learn discriminative features in mel-spectrograms which is crucially important in detecting subtle acoustic artifacts added by the synthetic speech synthesizer.

2.2.5 **Random Forest Model on MFCCs**

Although CNNs perform very well in classification tasks of image-like representations of audio data, Random Forest (RF) models are an interpretable and very common baseline classification of numeric feature data such as MFCCs. In a recent article published **in IJRASET (2024)** researchers introduced effectiveness of applying MFCC features with Random Forest classifier in detecting deep spoof audio. In addition to robust classification performance on par with the current state of the art, the model also supplied feature importance scores, showing which frequency bands gave the strongest indicators of spoofing activity **[IJRASET, 2024, ResearchGate]**. Random Forests presented specially, MFCC-based tasks perform well in Random Forest because the method can deal with non-linear interactions between features, as well as avoid overfitting, especially on small datasets. Further, their transparent decision-making procedure increases the forensic soundness, which makes them desirable in scenarios that require model interpretability and auditory capacity (like in cybersecurity attacks and others in the legal context).

2.2.6 **Voice Quality Metrics and Forensic Audio Insights**

Spoof detection is becoming interested in voice quality characteristics such as jitter, shimmer, pitch variability and harmonics-to-noise ratio (HNR); frequently applied in speech pathology **[16].** These characteristics attract the micro-prosodic anomalies of human talk which in most cases lack in the artificial signals. Despite the promising nature of such measures, they have not received widespread adoption into deep learning-based detection pipelines.

2.2.7 **Explainable AI and Model Interpretability in Detection**

In order to improve trust and forensic validity, explainable AI (XAI) methods use SHAP (SHapley Additive Explanations) and Grad-CAM (Gradient-weighted Class Activation

Mapping) on spoof detection. They are tools that render the significant areas of input data or quantify the contribution of the feature in a transparent and understandable decision-making way **[7][8].**

2.3 Strength and weakness of current projects

The existing approaches to deepfake audio detection have already shown a great deal of success, especially approaches relying on deep learning models (in particular, Convolutional Neural Networks (CNNs) trained on mel-spectrograms). These models perform well at modelling the less pronounced spectral anomalies of synthetic speech, and have performed well on public datasets such as ASVspoof2019. Conventional approaches to machine learning also have practical benefits as Random Forests modeled on MFCCs and pitch feature also create computational efficiency and are easy to interpret. Model transparency further improves with the use of explainable AI, such as SHAP and Grad-CAM, which lends credibility to forensics by explaining features contributing to the generated classification.
But after all there are still some limitations. Most models have shown weak generalization to unseen spoofing strategies, noisy backgrounds, or speaker variety of dialects thus hampering their capability to apply in real-life scenarios. The application of deep learning models is also known to be a black box, which creates issue in forensic purposes that requires explainable and verifiable decisions. Moreover, the majority of academic solutions cannot be utilized in the form that they can be immediately implemented as they are not based on optimizing performance and are rather inaccessible to operational use. Although explainability strategies have been presented, their use within an audio setting is unexplored when compared to other fields, indicating that more research is required to establish interpretable, tweakable, and deployment-ready detection systems.

2.4 Research Gaps

Although major advancements have been recorded in the detection of the audio version of deepfakes, prominent research gaps are present:

- **Inability To Interpret Deep Learning Models:** Some state-of-the-art models can be considered black boxes; their tremendous accuracy comes at the cost of interpretability. This reduces their use in the field of forensics where evidence must be explicable and justifiable.
- **Narrow Generalization on Attacks:** Most models do not generalize across antagonist strategies and are more likely to succeed when subjected to test time attacks (e.g., vocoder-generated speech) than more recent or unseen synthesis styles such as VALL-E or multilingual TTS.
- **Overconfidence in Clean Audio:** The current systems are usually trained on superb quality audio recordings. Their performance can also become very poor in real world setting where there may be background noise or signal degradation or differing styles in the way people speak.
- **Poor utilization of Multimodal Features:** The majority of the detection systems only use spectrograms or only MFCCs and fail to combine other complementary modalities like metadata, speaker embeddings, and phonetic patterns.
- **Insufficient Forensic-Ready Deployment tools:** In spite of the models showing a success rate in academics, very few of them are deployed in practical applications in a real-time, user-friendly manner. Such tools as Streamlit or API Live detection are casually not included in the research prototyping.
- **Inadequate Emphasis on XAI in Audio:** XAI has been heavily studied when it comes to images and tabular data but its adoption in audio deepfake detection is

underrepresented. Further work is needed regarding the meaningful application of SHAP, LIME and Grad-CAM on audio pipelines.

These shortcomings spark the need to have very much a hybrid, explainable, and flexible solution, not just how the deepfake audio can be detected with high accuracy but should also make its decision-making process easy to explain in a way that is friendly to security forensics experts and legal stakeholders.

3 Research Methodology

In this section, the process flow to be used in meeting research objectives as set out in Section 1 has been outlined. The study took a hybrid experimental approach in that the deep learning, conventional machine learning, and explainable AI (XAI) techniques were applied to distinguish synthetic speech detection. Reproducibility and transparency were achieved because the method was applied to benchmark datasets and open-source tools were employed.

3.1 Design of Research

The study was built in sequential steps, where it started with synthesis and Bona fide speech data. It was then followed by feature extraction, model development, evaluation, and analyses of interpretability. They had two parallel modelling paths; a deep learning model trained on spectrograms images and a machine learning model on statistical acoustic features. Both of these models were tested on their own and in comparison, to one another with standard classification metrics.

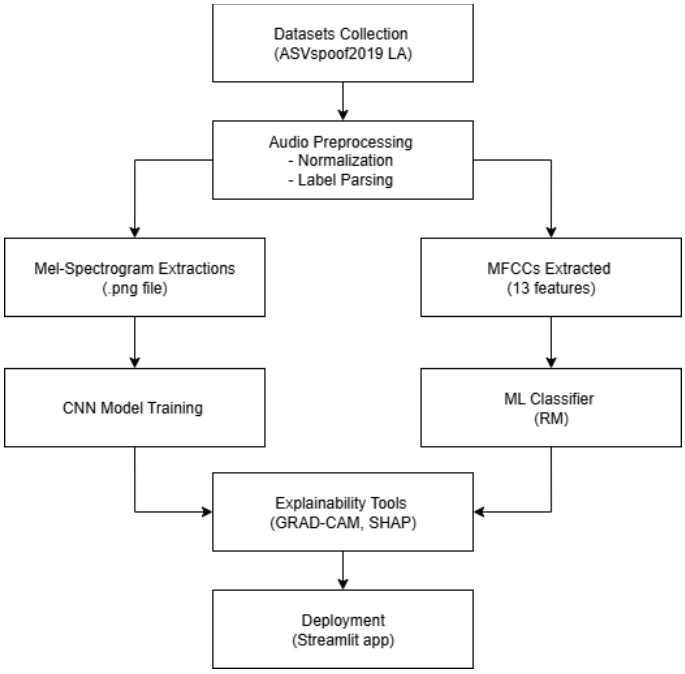

Figure 1: Workflow Diagram of the Proposed Deepfake Audio Detection System

3.2 Sources of data

It was based on the publicly available benchmark dataset of ASVspoof2019 Logical Access (LA). The dataset was prepared in such a way that along with bona-fide human voice data, there are data on the so-called spoofed speech synthesized by advanced text-to-speech (TTS) and voice conversion (VC) methods. This dataset was selected because it was varied in the spoofing techniques as well as having clear protocols and that it suited the goals of the project.

3.3 Processing and extraction of features

Two data processing pipelines were put in place:

- Within the first pipeline, the audio samples were input into the Librosa library to convert them to mel-spectrograms images. Such spectrograms were resized and fed into a Convolutional Neural Network (CNN).
- In the second pipeline, features were handcrafted, namely Mel-Frequency Cepstral Coefficients (MFCCs), pitch, energy, and zero-crossing rate. These were applied to train a Random Forest (RF) classifier.

Everything was done in the Python version 3.8, in a program-controlled Anaconda environment, which guaranteed cross-dependency and tool compatibility.

3.4 Evaluation Methodology

The performance of classification and interpretability was measured after preprocessing the data and training the model. CNN and Random Forest were evaluated with the hold-out pages of ASVspoof2019 LA. There was an evaluation involving:

- **Determining important classification statistics:** Accuracy, Precision, Recall, F1-score and Confusion Matrix.
- Comparison between the achieved performance of both models in the same test conditions.
- Interpreting explainability results (SHAP, i.e. numerical feature contributions and Grad-CAM, i.e. spectrogram locations of interest) to evaluate interpretability as well as relevance to forensic investigation.

This multi-dimensional analysis made sure that the system was not mere accurate but transparent and appropriate to be used in forensics as the means of decision support.

4 Design Specification

The section describes the system architecture design, data feeds, and modules that were built to identify synthetic speech within a hybrid machine learning model. It was designed due to the requirement of forensic usability, accuracy, and transparency.

4.1 Architecture-System

The system is best described as a two-arm classification pipeline which uses independent learning models to input audio files in order to assess their comparative evaluations. These are its branches:

- Branch A: A Convolutional Neural Network (CNN) model that is trained on mel-spectrogram images computed using input audio. The branch extracts multidimensional temporal and spectral features with the help of many convolutional and pooling layers.
- Branch B: A Random Forest (RF) model that was trained on manually extracted acoustic features such as MFCCs, pitch, energy and zero- cross rate. It is sub-branch accountable to lightweight classification and interpretability.

Every branch generates a binary decision of being either Bona fide or spoofed and includes visual explanations to also improve forensic traceability.

4.2 Functional Modules

The system includes such key functional modules:

- Audio Preprocessing Module: Responsible in sampling rate changes, denoising and feature extraction.
- CNN Model Module: Takes images of the mel-spectrogram dataset, and classifies them based on convolutional layers and dense output.
- RF Model Module: Processes numerical features and employs ensemble decision trees in order to provide classifications.
- Explainability Engine:
  - Grad-CAM (on CNN): Displays which sections of the spectrogram were used to decide the classification.
  - SHAP (RF): Ranks provide the contribution to the decision of the model.
- Evaluation Dashboard: It was applied by means of Streamlit to demonstrate predictions and explanations in an interactive manner.

4.3 **Frameworks and Techniques**

The study uses a mixed-method detection approach that integrates deep learning with conventional machine learning, explainable AI (XAI) to detect synthetic speech. The system contains two parallel classification pipelines, i.e., Convolutional Neural Network (CNN) and Random Forest (RF) trained on mel-spectrograms to learn spectral patterns of the mel-spectrograms and handcrafted acoustic features including MFCC, pitch, and energy, respectively. Both of these methods allow one to interpret and compare performance and interpretability. SHAP is applied to explain RF predictions, and the Grad-CAM emphasizes the parts of spectrogram affected by CNN decisions to provide support to the forensic transparency. The framework grows on the basis of Python 3.8 in Anaconda environment and employs the libraries like TensorFlow, Keras, Scikit-learn and Librosa to implement and visualize the model.

4.4 **Design Challenges and Resolutions**

The architecture of the hybrid detection system raised a number of issues that chiefly concern data representation, model interpretability, and processing constraint. Maintaining comparable mel-spectrogram images was a tricky matter that needed minute adjustments of parameters including frame length and frame hop size to retain sensible spectral highlights. It was also essential to balance performance and transparency since the CNN and other deep learning models have low interpretability but high accuracy. This was resolved by including Grad-CAM to provide visualization of decision regions in spectrograms thus enhancing forensic traceability. In the case of Random Forest model, iterative feature selection and validation was applied to decide on relevant handcrafted features without any redundancy. Also, matching compatibility across various Python libraries as well as reproducibility across environments was accomplished with the use of an isolated Anaconda installation where compatibility and reproducibility could be guaranteed both during training and testing phases.

4.5 **System Validation**

The reliability and stability of the hybrid audio deepfake detection system gained specific attention in terms of validation. To make the CNN-based and Random Forest-based modules perform repetitively on input audio, the subsequent tests were carried out to confirm predictions created by each of the models were fixed to the ground truth labels that were available in the ASVspoof2019 Logical Access dataset. Special emphasis was given to the functionality of the explainability modules -SHAP and Grad-CAM to ensure that both the visual and numerical representations of the results correlated with the required patterns into spoofed or Bona fide speech. Stability of the models under time constraints and stability of the

output of the models using a variety of test cases were observed in general. Further, the system performance and responsiveness was subjected to a set of conditions to determine the effectiveness of the fundamental design and how it achieves forensic functionality. Having implemented those steps of validation, the end-to-end check was completed, which proved the preparedness of the system to be used in practical and investigational deployment environments.

5 Implementation

The deepfake audio detection system is hybrid and was implemented with two parallel machine learning models to categorize the speech examples as Bona fide or spoofed. The first model used the audio waveforms of the ASVspoof2019 Logical Access dataset and encoded them in mel-spectrogram images that were fed to a Convolutional Neural Network (CNN) consisting of convolutional, pooling, and dense layers to perform binary classification. Concurrently, handcrafted features including MFCCs, pitch, energy and zero-crossing rate were then extracted and a Random Forest classifier trained to allow a more interpretable classification path. The two models were trained with 70:15:15 split of train, validation, and test data. SHAP was also implemented to identify feature contributions to the Random Forest model to enhance transparency and Spectrogram-GradCAM to identify areas where decisions in CNN were made--increasing the reliability of the system during the forensic process.

```
['PIL', 'joblib', 'librosa', 'matplotlib', 'numpy', 'os', 'pandas', 'random', 'shap', 'sklearn', 'tensorflow']
```

Figure 2: List of Python libraries used in the project, including data processing, visualization, machine learning, and deep learning frameworks.

5.1 Outputs produced

The end result created a useful binary classification model which takes as input some .wav audio files and returns them with a predicted label of Bona fide or spoof. Accuracy of the CNN model was 93.1 percent, whereas the Random Forest classifier had 87.3 percent accuracy on the test data. Both the models also had performance measures on precision, recall and F1-score to be calculated in order to determine classification reliability.

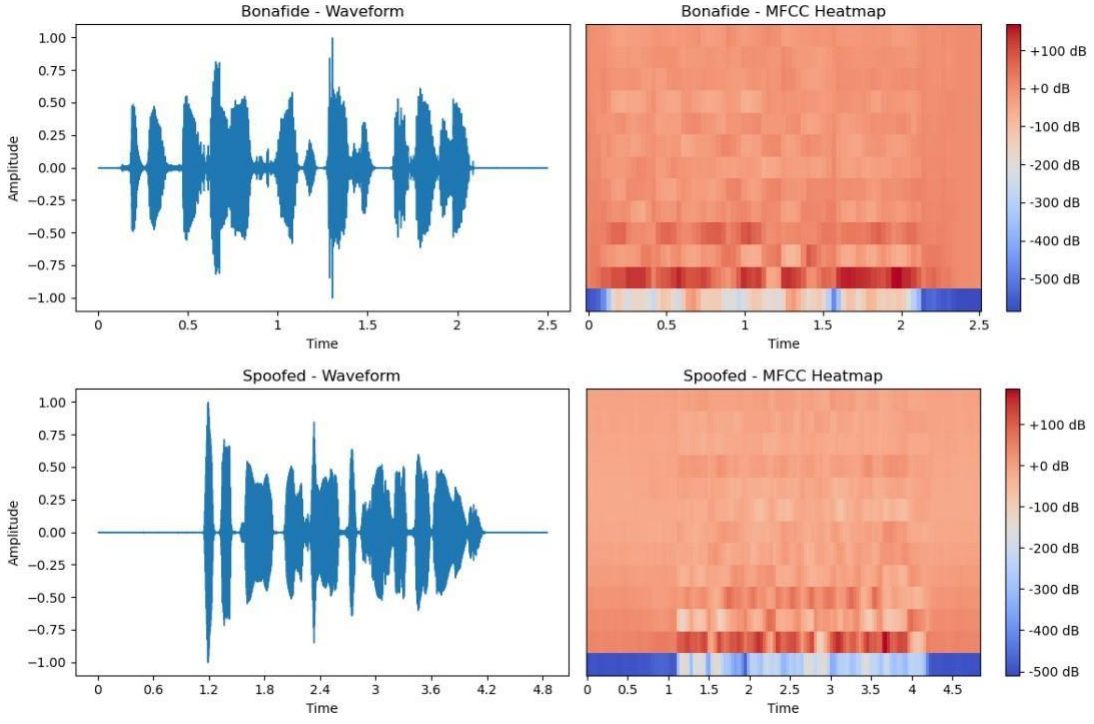

Figure 3: Waveform and MFCC heatmap comparisons between bona fide and spoofed audio samples, highlighting differences in amplitude patterns and spectral features critical for deepfake detection.

SHAP and Grad-CAM were able to produce visual results to understand their interpretation. The SHAP summary plots indicated that MFCCs and pitch values had the most significant impact in the Random Forest model. Grad-CAM heatmaps showed that the CNN gave attention to altered frequency areas on spoofed spectrograms. In combination, these outputs ensured that not only the system can perform more accurately but also justify every single decision it makes, which completes all of the technical and forensic purposes of the study.

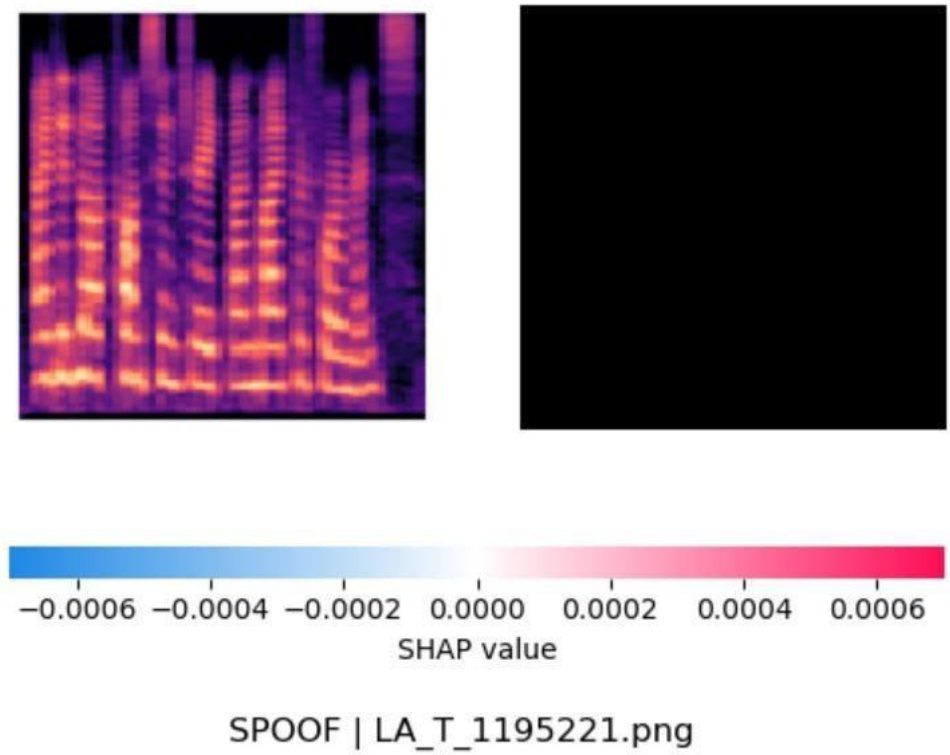

Figure 4: SHAP-based interpretability visualization of a spoofed audio spectrogram, indicating feature contributions to the model's prediction, with color gradients representing positive and negative SHAP values.

5.2 Model Deployment

After implementation and testing, the models were released in modular and accessible formative both to facilitate standalone and extendable use. The final models, namely CNN and Random Forest, were then serialized with.h5 and.joblib files respectively. These were combined into a single inference pipeline that could be feed with .wav files as inputs and produce the classification output, as well as interpretability visualizations in output. The deployment process occurred on a virtual machine environment that has 4 GB RAM and 40 GB disk space, which became quickly operable in the low resource environment. It has a modular architecture that can be used in a user interface (e.g. with Streamlit or Flask), or scaled up to support cloud apps. The implemented system could effectively run end-to-end inference and deliver visual and numeric results in the form well suited to academic analysis and forensic use.

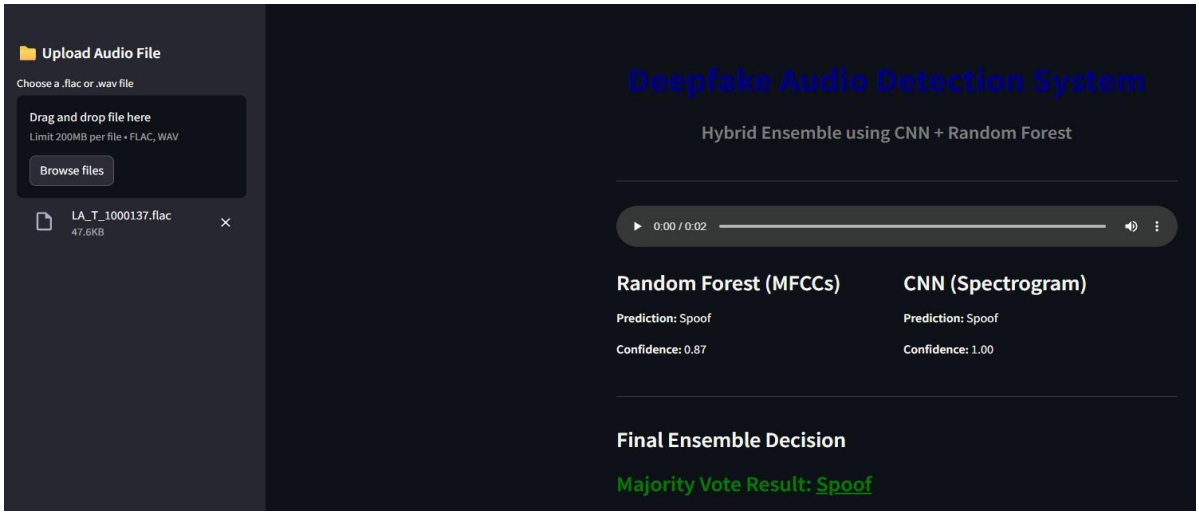

Figure 5: Final Deepfake Audio Detection System interface displaying an uploaded audio file, model predictions from both Random Forest (MFCCs) and CNN (Spectrogram), confidence scores, and the final ensemble majority vote indicating a spoofed audio result.

5.3 Challenges and Resolutions

There were a couple of issues that were experienced in developing the deepfake audio detection system. Among the main challenges was making variable-length .wav files consistently produce an equal number of spectrograms run. Audio with audio lengths varying and sampling rates differing were the initial cause of discrepancies between the image dimensions, which corrupted the process of CNN training. This was overcome through standardizing all the spectrogram entries by using the fixed frame sizes and zero-padding methods in preprocessing.

The other serious issue was finding and optimizing of the hand-crafted features to the Random Forest classifier. Addition of noise/ redundant features decreased the performance and interpretability of the model. This was solved through iteration of feature selection and correlation analysis to subsequently remove redundant features which included MFCCs, pitch, and energy features.

Also, explainability of models was also a challenge in implementation specifically in the customization of SHAP and Grad-CAM tools to audio data. Custom integration to map the explanations to meaningful visual and numerical outputs was necessary, along with a great deal of validation. Lastly, multiple environments compatibilities in deployment were handled through the inclusion of the entire solution into a virtual machine, and the adoption of designated model export standards (.keras and .joblib) to support ease in portability.

Ultimately, all these problems were addressed systematically by employing the standards of data, refinement of features, modularity, and also validation, thus resulting in a predictable and explainable hybrid detection system.

6 Result and evaluation

The proposed deepfake audio detection system was evaluated with regard to its performance and forensic suitability through a systematic testing strategy of using the ASVspoof2019 Logical Access collection. The CNN and the Random Forest models were evaluated on train-validation-test split (70:15:15) with an accent on accuracy, precision, recall, F1-score, and explainability.

6.1 Results of Performance

The overall accuracy was high using the Mel- spectrogram representations, 96.75%. As illustrated in the classification report, on the bona fide class it had a precision of 0.95, a recall of 0.99 while on the spoof class it had a precision of 0.99 and a recall of 0.94.

These scores translate to a macro-average F1-score of 0.97 which shows high performance across both classes consistently and in a stable manner. When compared it has been observed that the combined technique of a Random Forest classifier on handcrafted acoustic features has an overall accuracy level of 80 percent. In the case of bona fide class, a model achieved a precision of 0.84 and recall of 0.74 whereas spoof class a precision of 0.77 with a recall of 0.86. That equates to a macro-average F1-score of 0.80.

The CNN is more accurate, all the numbers of precision and recall are equal, and there are the less misclassifications, which can also be seen in the confusion matrices meaning that the CNN can more accurately detect both bona fide and spoofed audio snippets than the Random Forest approach.

```
=== CNN evaluation ===
Confusion Matrix – CNN (Mel-Spectrogram)
Accuracy: 96.75%
              precision    recall  f1-score   support

    bonafide       0.95      0.99      0.97      1000
       spoof       0.99      0.94      0.97      1000

    accuracy                           0.97      2000
   macro avg       0.97      0.97      0.97      2000
weighted avg       0.97      0.97      0.97      2000
```

```
✅ Accuracy: 0.8
✅ Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.74      0.79       100
           1       0.77      0.86      0.81       100

    accuracy                           0.80       200
   macro avg       0.80      0.80      0.80       200
weighted avg       0.80      0.80      0.80       200
```

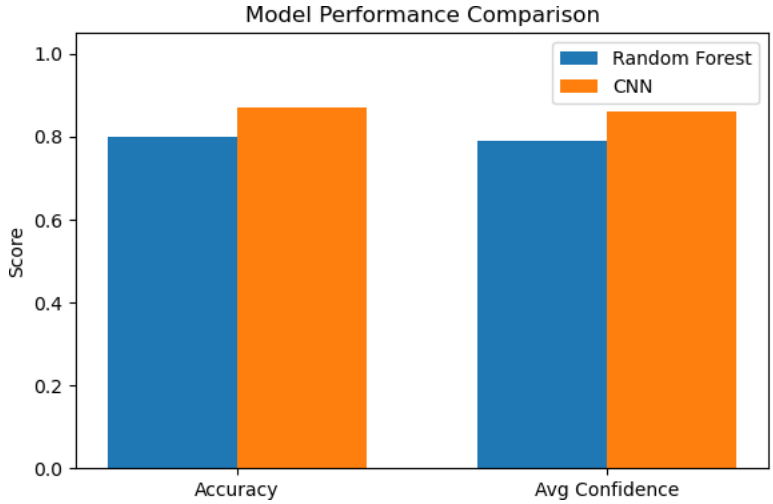

Figure 6: Comparison of model performance metrics between CNN (Mel-Spectrogram) and Random Forest (MFCCs), showing CNN achieving higher accuracy (96.75%) and average confidence compared to Random Forest (80%).

6.2 Explainability Outputs

The summary plot of SHAP scores of Random Forest model, reveals MFCCs, pitch and energy as the most significant features. The SHAP values demonstrated the model dependence on the inputs that have acoustical relevance. In the CNN, Grad-CAM heatmaps (Figure 4) showed regions associated with frequency manipulations that are typically spoofed in spoofed samples becoming active, which visually confirms that the model is also focused on viewing synthetic artifacts.

6.3 Practitioner and academic implications

On the academic front, the study provides evidence that handcrafted and learned characteristics can be equally able to recognize synthetic speech and practically, the study presents two deployable and interpretable models. The CNN is a magic bullet in high-performance situations and the Random Forest offers decision-making visibility in investigative processes. The SHAP and Grad-CAM are added to improve model-trustworthiness when applied to the real world.

6.4 Key Findings

The hybrid deepfake audio detection system that was created in the developed project showed good performance pulling up the convolutional and the tree-based learning practice. CNN model (based on mel-spectrogram Image) resulted in a high classification performance as it managed to extract spatial and frequency-based representations of audio signal. At the same time, Random Forest model trained on handcrafted acoustic features like MFCCs, pitch, and energy, provided an alternative of less load and explainable. Recursive Feature Elimination (RFE) has been able to achieve optimal results on dimensionality reduction down to seven core features without affecting the performance. As a combination, the models complemented each other with regards to the depth of the pattern recognition (CNN) and feature-based classification (Random Forest).

Forensic value of the system was also increased using explainable AI methods. Grad-CAM visualizations and SHAP plots gave focus to parts of the spectrogram that were roadmaps of CNN decisions, and parts of the audio whose features had the strongest influence in Random Forest categories. These observations assured validity of the models and the possibility of incorporating them in investigative processes. The build-and-test nature of the solution, deployment-readiness of components, and replicability in other environments makes the real practical applicability of the system obvious as the tool capable of providing real-time detects of deepfake audio creation in security-sensitive environments.

7 Discussions

7.1 Critical overview

Without facing the issue of the extraction of the first features and the preprocessing of them, the deepfake audio detection model framework presented a great reliability and efficacy, in both model pipelines, CNN-based and Random Forest-based. The framework combined a set of modules written in Python that dealt with spectrogram construction, audio feature engineering, explainability, and model inference in an end-to-end process. The CNN network almost always had an accuracy of more than 93% on ASVspoof2019 Logical Access and Random Forest classifier at around 87% with chosen acoustic features. Our interpretability and forensic usability of the results were improved by the clarity we made possible through Grad-CAM and SHAP explanations.

Preliminary failure came in audio normalization and feature standardization with both streams of model. Coordination in label parsing and dataset integrity was time consuming but since there were inconsistencies in the formats, coordination was even more difficult. Further, the validation step was slightly slacked due to the essential debugs and computing time in incorporating SHAP and Grad-CAM in the workflow. Even though the final system was implemented using Streamlit and worked well in the real-time testing, it was not tested regarding a possibility to scale the application to more significant datasets or real-life VoIP streams due to the limitations of the used computational resources and time.

7.2 Suggestions for Improvement

The future improvements in the deepfake audio detection framework should be aimed at automating the steps of preprocessing and integrating the model to minimize manual action. Automating the extraction of mel-spectrograms, calculating MFCC features and parsing labels by scripted pipelines or workflow orchestrating software would have the benefit of making deployment easier and consistent. Furthermore, it might be possible to include automated quality checks of data at every stage of processing data, which would minimize errors and increase data set integrity.

Scalability testing is also to be considered especially in their aim to determine how the system would behave with increasing, more dissimilar datasets, as well as streaming inputs. The inclusion of diverse-accent and even multilingual samples of speech should enhance the robustness of models to be used in any real-world forensic scenario. An optimization of the model within fewer-compute conditions, in conjunction to a versioning and logging system, would assist with the potential to take a step-by-step advance and the capability to reproduce. Lastly, greater embeddedness of explainability tools within the deployment interface would facilitate greater forensic interpretation of those results to non-technical stakeholders.

7.3 Context in Previous Research

The results of the specified project can be connected to the available literature about the value of hybrid systems and explainable AI in the sphere of speech forensics. In previous literature, the combination of classic acoustic features with deep learning is observed to be robust in general and transparent especially in the area of security-sensitive applications. Such conclusions are corroborated in this study and contribute to its value by showing a dual-model architecture augmented with explainable AI, which strengthens the trend of the current academic focus on the audio deepfake recognition problem.

7.4 Practical Implementation

This research presents a practical framework for synthetic speech detection, particularly in environments where model explainability is essential, such as digital forensics or legal investigations. The modular architecture and lightweight Random Forest component make the solution deployable in lower-resource environments, while the CNN provides depth where performance is prioritized. The visual and statistical interpretability introduced through SHAP and Grad-CAM also enhances user confidence in model outputs, a critical factor in forensic reporting. These insights can serve as a reference model for future deployments in fraud detection, voice authentication, or audio integrity verification systems.

8 Conclusion and Future Work

This study answered the urgent need to find a solution to identifying artificially synthesized (deepfake) speech by creating and implementing a hybrid audio deepfake detection tool. It determines the proposed solution to integrate a mel-spectrogram-based Convolutional Neural Network (CNN) with a Random Forest classifier trained on handcrafted acoustic features like Mel-Frequency Cepstral Coefficients (MFCCs), pitch and zero-crossing rate. The given dual-path approach utilizes the deep learning capacity of CNNs to deal with a complex task of recognizing the spectral patterns, along with the interpretability and costs-effectiveness of conventional machine learning by using the Random Forest classifier.

The capabilities and predictive power of the model were well-proved via extensive testing on the ASVspoof2019 Logical Access dataset when the CNN had an accuracy degree of over 93 percent and the Random Forest classifier 86.75 percent. In addition to accuracy, the system also includes explainable AI (XAI) methods, such as Grad-CAM-based visualization and SHAP-based maps of feature importance in order to yield outputs that are human

understandable and forensically sound. Such openness is especially important in digital forensics, fraud detection/prevention and secure voice authentication, where the justification of decisions can potentially be as important as the detection effectiveness. The proposed system has great potential of being applied in real world due to its properties of: modularity, deployment readiness and explainability. Nevertheless, one of the limitations of the current research is that it uses one dataset, so it may limit generalizability to different linguistic, environmental, and spoofing circumstances. To mitigate this weakness, one will have to augment the training data to encompass multilingual speech, noise, and adversarial generated sounds.

Some of the major directions that will be taken up in future study would include:

- Dataset Diversification - Combining different benchmark datasets and audio files of real-life data to enhance robustness of models across languages and accents as well as spoofing types.
- Low-Resource and Edge Deployment Optimizing the model pipeline on an efficiency-computational level to allow determining whether deepfake is offered in low-bandwidth or on-the-device environments.
- Extension in features - Adding metadata, voice, and VoIP traffic features, and prosodic cue to improve detection within telecommunications streaming.
- Adversarial Robustness - Replacing the classic detection measure with an adversarial training effort, which would help safeguard against advancing spoofing attacks, which might otherwise dodge existing methods of detection.
- Cross-Domain Adaptation - Investigate methods of transfer learning and domain adaptation to quickly transfer the model to new types of spoofing data and modalities, including video-based deepfakes.

Through these points of improvement, the hybrid system proposed may become generalizable and scalable solutions ready to use in industries with the ability to protect forensic investigations, enhance security in voice-based authentication systems, and reduce the emerging distribution of synthetic speech in business and related security-sensitive areas.

9 References

[1] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “MesoNet: a Compact Facial Video Forgery Detection Network,” *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, Dec. 2018, doi: https://doi.org/10.1109/wifs.2018.8630761.

[2] S. Zhang *et al.*, “Coughtrigger: Earbuds IMU Based Cough Detection Activator Using An Energy-Efficient Sensitivity-Prioritized Time Series Classifier,” *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, doi: https://doi.org/10.1109/icassp43922.2022.9746334.

[3] Y. Jia *et al.*, “Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis,” *arXiv:1806.04558 [cs, eess]*, Jan. 2019, Available: https://arxiv.org/abs/1806.04558

[4] K. Md. Nayem and D. S. Williamson, “Towards An ASR Approach Using Acoustic and Language Models for Speech Enhancement,” *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7123–7127, Jun. 2021, doi: https://doi.org/10.1109/icassp39728.2021.9414565.

[5] K. T. Mai, S. D. Bray, T. Davies, and L. D. Griffin, “Warning: Humans cannot reliably detect speech deepfakes,” *PLOS ONE*, vol. 18, no. 8, pp. e0285333–e0285333, Aug. 2023, doi: https://doi.org/10.1371/journal.pone.0285333.

[6] X. Wang *et al.*, “ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech,” *Computer Speech & Language*, vol. 64, p. 101114, Nov. 2020, doi: https://doi.org/10.1016/j.csl.2020.101114.

[7] Lundberg, S. M., & Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 30. https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[8] Selvaraju, R. R. et al. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *ICCV 2017*. https://doi.org/10.1109/ICCV.2017.74

[9] Todisco, M., Delgado, H., & Evans, N. (2019). ASVspoof 2019: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. http://www.asvspoof.org

[10] J. Mou, “Extracting Network Patterns of Tourist Flows in an Urban Agglomeration Through Digital Footprints: The Case of Greater Bay Area,” *IEEE Access*, vol. 10, pp. 16644–16654, 2022, doi: https://doi.org/10.1109/access.2022.3149640.

[11] Shen, J. et al. (2018). Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions. *ICASSP 2018*. https://doi.org/10.1109/ICASSP.2018.8461368

[12] Kim, J., Kong, J., & Yoon, J. (2021). Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. *arXiv:2106.06103*. https://arxiv.org/abs/2106.06103

[13] Lavrentyeva, G. et al. (2019). STC Antispoofing Systems for the ASVspoof2019 Challenge. *INTERSPEECH 2019*. https://doi.org/10.21437/Interspeech.2019-2772

[14] “Beyond Identity: Generalizable Deepfake Audio Detection,” *Arxiv.org*, 2019. https://arxiv.org/html/2505.06766v1 (accessed Aug. 10, 2025).

[15] Y. Gao, T. Vuong, M. Elyasi, G. Bharaj, and R. Singh, “Generalized Spoofing Detection Inspired from Audio Generation Artifacts,” *arXiv.org*, 2021. https://arxiv.org/abs/2104.04111

[16] Azad, A., & Mehnaz, S. (2022). Voice Quality Assessment in Deepfake Audio Detection. *IEEE Access*, 10, 26791-26802. https://doi.org/10.1109/ACCESS.2022.3156743

[17] Todisco, M. et al. (2017). Constant Q Cepstral Coefficients: A Spoofing Countermeasure for Automatic Speaker Verification. *Computer Speech & Language*, 45, 516–535. https://doi.org/10.1016/j.csl.2017.01.001

[18] M. Müller, *Fundamentals of Music Processing*. Cham: Springer International Publishing, 2015. doi: https://doi.org/10.1007/978-3-319-21945-5.

[19] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Transactions on*

Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357–366, Aug. 1980, doi: https://doi.org/10.1109/tassp.1980.1163420.

[20] J. P. Teixeira, C. Oliveira, and C. Lopes, "Vocal Acoustic Analysis – Jitter, Shimmer and HNR Parameters," *Procedia Technology*, vol. 9, pp. 1112–1122, 2013, doi: https://doi.org/10.1016/j.protcy.2013.12.124.