gradDoc

by 211867 Zakaria Sameh Elimam Attia Omar

Submission date: 23-Jun-2024 07:55PM (UTC+0200)
Submission ID: 2406697004
File name: 241709_211867_Zakaria_Sameh_Elimam_Attia_Omar_gradDoc_318893_1224022677.pdf (1.25M)
Word count: 9733
Character count: 57156

Faculty of Computer Science

Proposal Document for project Audio deep-fake detector for banking systems

Zakaria Sameh 211867
Supervised by: Dr. Ali Hamdi

June 23, 2024

Contents

List of Figures

Abstract

In the era of rapid advancements in artificial intelligence, the ability to manipulate audio to create deep fakes has reached a level of sophistication where the human ear can no longer reliably distinguish between authentic and synthetic speech. This technological prowess, while beneficial in various applications such as entertainment and accessibility, has also given rise to a plethora of malicious uses, particularly in the realm of financial fraud. The prevalence of such manipulated audio in criminal activities underscores the urgent need for robust systems capable of identifying spoofed audios across a wide range of scenarios.

This project proposes the development of an advanced system designed to combat the threat of audio deep fakes within the critical domain of banking systems. The system is envisioned to serve as a pivotal component in the authentication process, ensuring the integrity and security of transactions conducted through voice-driven devices (VDDs). These devices, which rely on automatic speaker verification (ASV) systems, are inherently susceptible to various types of attacks, including voice-based logical access (LA) and physical access (PA) attacks.

The core of the proposed system architecture is a speaker verification model that utilizes Long Short-Term Memory (LSTM) networks, a type of recurrent neural network well-suited for processing sequential data such as speech. This model is complemented by a speaker identification model, both of which leverage Mel-Frequency Cepstral Coefficients (MFCCs) as key features for distinguishing between genuine and spoofed audio. MFCCs are chosen for their effectiveness in capturing the spectral properties of speech, which are crucial for the accurate identification and verification of speakers.

The system's design is grounded in the need for a comprehensive solution that can not only detect current spoofing techniques but also adapt to emerging threats. To achieve this, the system will be trained and tested on a diverse set of datasets, including those from the ASVspoof challenges, to ensure its robustness and reliability. The integration of the LSTM model with MFCC features represents a significant step towards creating a secure and efficient authentication system that can withstand the evolving landscape of audio deep fake technology.

The project's ultimate goal is to deliver a system that can be seamlessly integrated into existing banking infrastructure, providing an additional layer of security against fraudulent activities. By doing so, the system aims to protect both the financial institutions and their customers from the potential risks associated with audio deep fakes, ultimately contributing to the enhancement of trust and confidence in voice-driven banking services.

In summary, this project proposes the development of a cutting-edge audio deep fake detection system tailored for the banking sector. The system's architecture, which combines LSTM networks with MFCC feature extraction, is designed to provide a reliable and scalable solution for identifying spoofed audios. Through rigorous testing and continuous adaptation, the system aims to stay ahead of the curve in the ongoing battle against audio deep fake technology, ensuring the safety and authenticity of voice-driven transactions in the banking industry.

CH 1 : Introduction

AI-manipulated and generated audios are the resulting audios created through a machine instead of a human reciter, their general purpose was automating a lot of tasks like producing audiobooks instead of wasting time and resources on transcriptions, also entertainment, creating music, etc.
As for being used in entertainment, it causes the industry to move towards clearer, less robotic voices making it as realistic as it can get and the massive amounts of voice recordings that are broadcasted daily providing the material for more advances makes the need for an easy and equally accessible fake audio detector. Nevertheless, these methods are also used in harmful ways as impersonation which (depending on the target) can be used to manipulate public opinion for propaganda, defamation, and spreading misinformation, but this work's concern is mainly fraud.

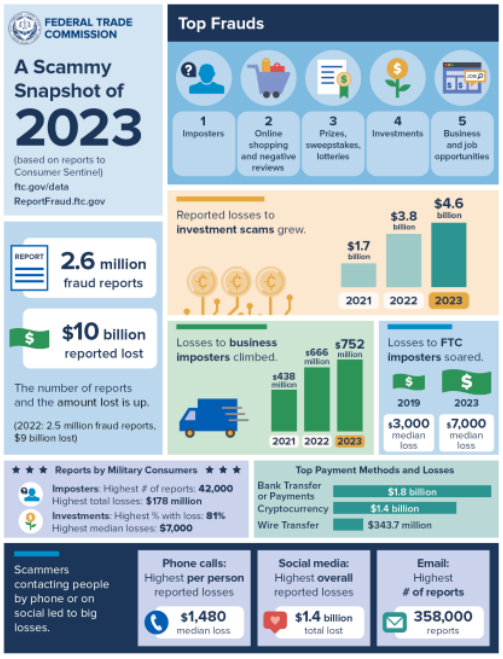


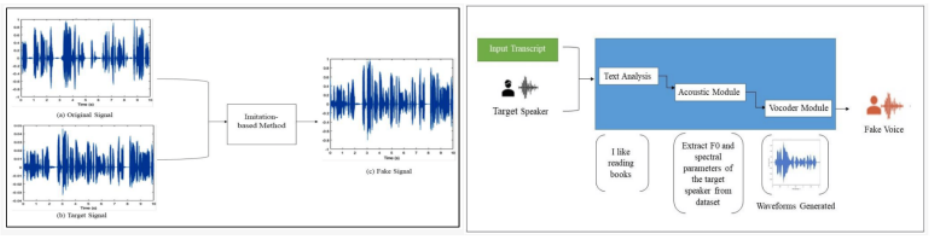
Figure 1: Federal Trade Commission report February 9, 2024

Fraud from banks and investment firms perspective is a leading and growing danger as consumers reported losing more money to investment scams (more than 4.6 billion) than any other category in 2023. That amount represents a 21 increase over 2022. The second highest reported loss amount came from imposter scams, with losses of nearly 2.7 billion reported. In 2023, Consumer Reports says the business sector will cost more than any other category in 2023 (more than $4.6 billion)as shown in figure (1). Instead of fraudlosses of approximately 2.7 billion were reported. In 2023, Consumers reported losing more money through wire transfers and cryptocurrencies than all other methods combined (It's also worth noting that these numbers represent the US only).

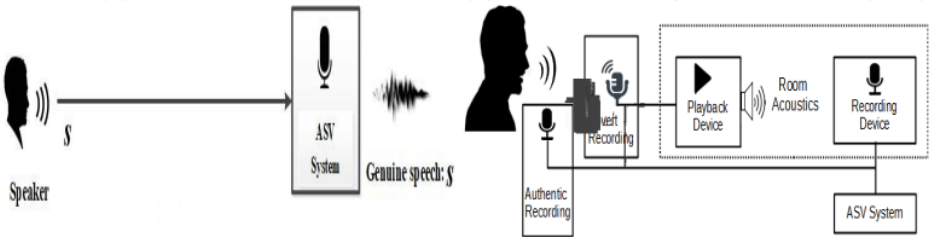
While these losses have huge financial impact on victims. It also damages a customers' credit score making it difficult to get loans or other forms of credit in the future, also a big impact on these entities reputation.

1.1 Background

The field of artificial intelligence has seen rapid advancements, particularly in deep fake voices, enabling the creation of indistinguishable fake voices. This progress has led to the proliferation of realistic DeepFakes. however, deepfakes are not the main concern but the methods they are used in and the attacks overall. There are four principal means to generate attacks: impersonation, voice conversion (VC), text-to-speech synthesis (TTS), and replay based.



(a) voice conversion (VC) (b) Synthetic-based / Text-To-Speech (TTS)



(c) Imitation-based (d) Replay-based

Figure 2: Types of attacks

As simplified in Figure 2, attacks on voice authentication systems can be categorized into four primary forms, each with its own unique characteristics and challenges for detection and prevention.

a) Voice Conversion (VC): This method involves an attacker obtaining two audio samples—one with the target's voice and another with the speech to be mimicked. The attacker then uses software to convert the speech to match the target's voice characteristics. This technique is particularly challenging because it requires the system to detect subtle alterations in the voice that may not be immediately apparent to human listeners. The converted audio may retain the linguistic content and style of the original speaker while adopting the vocal identity of the target, making it a sophisticated form of impersonation.

b) Synthetic-based / Text-To-Speech (TTS): In this approach, the attacker generates audio from scratch. The process begins with the attacker learning the target's

voice characteristics, often through machine learning algorithms that analyze a large corpus of the target's speech. Once the system has learned the voice, it can synthesize speech based on input text. This method is powerful because it allows the creation of entirely new utterances that the target has never spoken, potentially leading to the issuance of unauthorized commands or the falsification of consent.

c) Imitation-based: This category has historically received less attention compared to other methods, partly because it does not involve spoofing in the traditional sense. Instead, it deals with genuine speakers who may be impersonating another individual. This is not a voice verification problem per se but rather an identification challenge. The system must be capable of distinguishing between individuals who may sound similar or who intentionally alter their voice to mimic another person. This form of attack is particularly difficult to detect because it relies on human ability and intent rather than technological manipulation.

d) Replay-based: This is a straightforward yet effective method where an attacker either splices together bits of previously recorded audio to form a targeted speech or simply replays a pre-recorded audio file. While this method may seem rudimentary compared to the sophisticated techniques of voice conversion and TTS, it can be surprisingly effective, especially in scenarios where the authentication system does not have adequate protection against replay attacks. The challenge lies in detecting the replayed audio in real-time and distinguishing it from a live speaker.

Each of these attack methods presents distinct challenges to the security of voice authentication systems. Voice conversion and synthetic-based attacks require the system to identify subtle cues that differentiate genuine speech from manipulated audio. Imitation-based attacks demand a high level of discrimination to differentiate between individuals with similar vocal characteristics or those who are intentionally altering their voice. Replay attacks necessitate the system to be vigilant against pre-recorded audio that may be indistinguishable from live speech under normal circumstances.

To combat these threats, the proposed system architecture in this project incorporates advanced machine learning models, such as LSTM networks trained on MFCC features, to detect anomalies in voice patterns that may indicate an attack. The system is designed to be adaptable and robust, capable of learning from new types of attacks and improving its detection capabilities over time. By understanding the nuances of each attack method, the system can be fine-tuned to provide a higher level of security against the ever-evolving landscape of voice fraud.

In conclusion, the four forms of attacks—voice conversion, synthetic-based, imitation-based, and replay-based—each pose unique challenges to the security of voice authentication systems. The proposed system in this project is designed to address these challenges through the use of advanced machine learning techniques and a comprehensive understanding of the various methods used by attackers. By continuously improving and adapting to new threats, the system aims to ensure the integrity and security of voice-driven transactions in the banking sector and beyond.

1.2 Research Challenges

Detecting attacks on Automatic Speaker Verification systems (ASVs) is a complex task due to the diverse range of manipulation methods and scenarios encountered in real-world applications. The variability in bonafide audios further complicates the challenge, leading to several key issues and challenges:

1.2.1 Excessive Processing and Preprocessing

Traditional ASV systems often rely on Mel-Frequency Cepstral Coefficients (MFCCs) or pitch information for feature extraction. While these methods are effective for genuine voice verification, they require extensive processing and preprocessing to detect sophisticated voice-based attacks, such as deepfakes. The need for detailed feature engineering and the computational overhead associated with it can lead to delays in real-time applications, impacting the usability and efficiency of the ASV systems.

1.2.2 Balancing Between All Models and Their Trade-offs

The development of ASV systems involves the selection and optimization of various machine learning models. Each model has its own set of trade-offs, such as the balance between accuracy and computational efficiency. For instance, deep learning models like LSTMs and transformers can capture complex patterns in audio data but may require significant computational resources. Conversely, simpler models might be more resource-efficient but could lack the necessary complexity to detect advanced spoofing techniques. Finding the right balance between model complexity, accuracy, and resource consumption is a critical challenge in ASV system design.

1.2.3 Generalization to New Techniques

As audio manipulation techniques continue to evolve, ASV systems must be capable of generalizing to new types of attacks that were not present during the system's training phase. This requires the system to have robust feature extraction methods and learning algorithms that can adapt to emerging threats. The current reliance on specific feature extractors may limit the system's ability to detect novel attacks that do not conform to the patterns learned during training. Therefore, there is a need for ASV systems to incorporate more flexible and adaptable methods, such as attention mechanisms and transformers, which have shown promise in other domains but are yet to be fully exploited in audio deepfake detection.

In summary, the challenges faced by ASV systems in detecting sophisticated attacks are multifaceted, involving issues related to processing and preprocessing demands, the trade-offs between different models, and the ability to generalize to new spoofing techniques. Addressing these challenges is crucial for the development of robust and efficient ASV systems that can effectively protect against the growing threat of audio deepfakes in real-world applications.

1.3 Research Problem Statement

- Research often relies on observational studies and MFCC-based speech representation, which may not be effective against sophisticated voice-based attacks.
- Other methods, such as those using autoregressive models or convolutional neural networks, may struggle with data that does not follow the patterns seen during training.
- These approaches can lead to poor generalization in real-world scenarios and are challenged by the evolving nature of audio manipulation techniques.

1.4 Research Aim

The primary aim of this research is to develop a robust system capable of identifying spoofed audios across a diverse range of manipulation methods and scenarios encountered in real-world settings. Specifically, the research aims to address the pressing need for effective detection mechanisms within banking authentication systems, where the ramifications of fraudulent activities can be particularly severe.

By leveraging advances in machine learning and signal processing techniques, the research seeks to design and implement a comprehensive solution that not only identifies manipulated audios but also distinguishes them from genuine recordings with high accuracy. This entails developing novel algorithms and methodologies that can effectively capture subtle artifacts and inconsistencies indicative of spoofing.

Furthermore, the research aims to create a versatile framework that can adapt to emerging threats and evolving manipulation techniques. This involves not only achieving high detection rates on known types of attacks but also anticipating and mitigating vulnerabilities to novel forms of audio manipulation.

1.5 Research Methodology

The research objectives are outlined as follows: Data Collection: Assemble a diverse dataset encompassing genuine and spoofed audio samples across various manipulation methods and scenarios. Feature Discrimination: Develop techniques to extract discriminative features from audio signals to differentiate between genuine and spoofed recordings effectively. Model Development: Design and optimize deep learning models specifically tailored for the task of spoofed audio detection, aiming for high accuracy and robustness. Evaluation and Validation: Rigorously evaluate the performance of developed models through comprehensive testing against diverse manipulation methods and variations in audio quality. Generalization and Adaptation: Investigate methods to enhance the generalization capabilities of the detection system, enabling it to detect unseen manipulation methods and adapt to evolving threats. Integration and Deployment: Integrate the developed detection system seamlessly into banking authentication frameworks, ensuring compatibility and minimal disruption to existing systems.

1.6 Research Scope

The scope of this research focuses on the detection of audio deep fakes in the context of English language using simple evaluation metrics, primarily accuracy.

Focus Areas

- Audio Deep Fakes: The research investigates methods and techniques specifically tailored for detecting deep fakes in audio recordings.
- English Language: Emphasis is placed on detecting deep fakes in spoken English, considering linguistic characteristics and variations.

1.7 Research Significance

This research is significance is in it addressing a critical need for robust security measures in banking systems. By developing a system capable of detecting fake audio and identifying speakers, it can help prevent fraud and protect sensitive financial information. Additionally, the research has broader implications for improving security in other voice-driven applications beyond banking, contributing to advancements in the field of audio authentication technology.

1.8 Research contributions

This research's contributions are evaluating and testing various models to identify the most effective ones for detecting fake audio and identifying speakers in banking security systems. Additionally, it provides a functional pipeline and a user-friendly interface for the system, enhancing its usability and accessibility. Furthermore, the research involves updating and balancing the dataset to ensure its relevance and representativeness, thereby improving the accuracy and reliability of the system.

2.1 Overview

Compared to visual DeepFakes, voice detection DeepFakes is a new task with less resources, detection methods and good results. ASVspoof is one of the most important problems encountered in spoof detection. The 2021 version of the challenge brings a new one; In addition to Logical Access (LA) and Physical Access (PA), there is now a subset of DeepFake (DF). The goal is a solution that detects processed, compressed audio files commonly encountered on the internet. Although compression effects can be seen in some search solutions, they cannot be generalized across different datasets. It is worth noting that the purpose of the DF subset is only to generate speech and its main purpose is not to deceive the ASV system. The aim of their performance on the new DF task is to evaluate the robustness of false positive solutions in the search of speech materials published online with different properties.

The entire Deep Fake evaluation dataset was built using over 100 different spoofing attack algorithms developed by VCC participants as well as various groups and individuals who contributed to the ASVspoof 2019 competition. The results on the df showed vulnerability to lower quality and unseen fake audio. Additionally, detection models struggled against partially-spoofed audio.
The following study pointed out the poor performance validation and testing results of detecting false audios. Although, deep learning approaches achieves better results in detecting deepfake content, they often require extensive training time and computational resources. To work their way handling the limitations of higher feature sets and complexities, the study proceeded to lean on transfer learning to adapt to desired tasks to shift their focus to relevant features such as Mel-frequency cepstral coefficients (MFCCs).

MFCCs are mentioned in any human speaker related tasks, that are widely used in speech recognition tasks due to their ability to capture the decisive characteristics of the human voice, and can by then be used either for identification as the model learns to differentiate between voice biometrics such as pitch, accent, speed of speech, cadence, and tone, or for validation as the models focus on the qualitative attributes of these features. Nonetheless, mfccs are The primary reference.
The MFCC method involves several steps, including framing the audio signal into short segments, applying a window function to each frame, computing the discrete Fourier transform (DFT) of the signal, and finally calculating the MFCC coefficients using a filter bank that mimics the human auditory system's frequency response, but a notable mark is that these coefficients being a frequency domain features, do not serve as good for specific models for being unsuitable as their best performing features indicated. Spectral features that were used are roll-off point, centroid, contrast, and bandwidth were used alongside MFCCs to achieve an accuracy 8 to 23 percent higher that MFCCs along and other tested methods, so for our project we adapted mel-spec features too.

The roll-off point also represents the frequency which a certain percentage of the total signal energy is contained, it measures the spectral shape of the audio signal and can provide insights into the energy distribution across different frequency

bands. In deepfake detection , the roll-off point features' role is to help characterize the frequency content of the audio signal and identify potential anomalies or abnormalities that may indicate manipulation. The centroid is a measure of the center of mass of the frequency distribution, meaning it is an indication of the average frequency of the signal, it provides information about the spectral balance and the tones characteristics. By analyzing the centroid feature, we gain insights into the overall spectral profile of the audio signal and detect variations that may arise from deepfake manipulation. Contrast refers to the difference in energy between peaks and valleys in the spectral envelope, think of it as the model sees and handles as we think of variance and how is it used, it helps in analysis and noticing changes in the spectral dynamics of the audio signal. Bandwidth can identify variations in the frequency distribution as well, all these different indicators are used to show different perspectives of the desired area in the audio signal that is worth noting helping the model to adapt and focus on said attributes. Nevertheless, these studies still struggles against adversarial noise attacks and real world noises, on top of not being frail and vulnerable over dynamic patterns and across different attack types.

2.2 Taxonomy

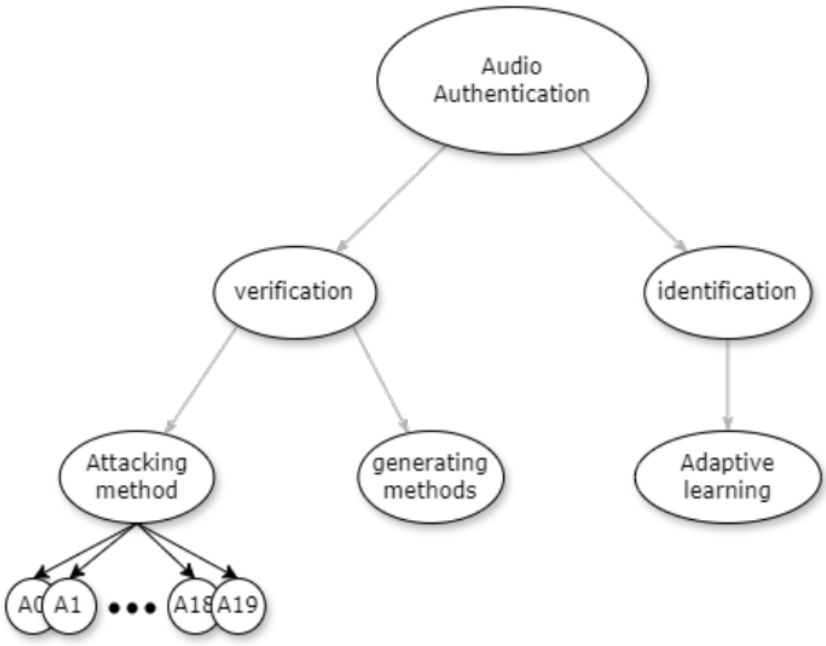


Figure 3: Taxonomy

This section categorizes and organizes the different types of audio manipulation methods, particularly focusing on those defined in the ASVspoof Challenge, which are crucial for banking security systems. As shown in figure (3) the ASVspoof Challenge introduces a taxonomy of attack methods labeled A0 to A19 [**5**], which encompass a range of techniques used to create fake audio. A0: This represents

genuine audio with no manipulation.
A1 to A4: These methods involve voice conversion techniques, where the attacker's voice is converted to sound like the target speaker's voice.
A5 to A8: These are text-to-speech synthesis methods, generating speech directly from text inputs to mimic the target speaker's voice.
A9 to A12: These methods use replay attacks, where a recorded genuine speech is played back to the system.
A13 to A16: These methods include more advanced synthetic speech attacks that combine multiple techniques to improve the realism and effectiveness of the spoofing.
A17 to A19: These are adversarial attacks that introduce subtle perturbations to audio samples to fool the detection systems while remaining imperceptible to human listeners.
By organizing these methods into distinct categories, the taxonomy facilitates systematic analysis and comparison, enabling researchers to identify common patterns and trends in audio spoofing techniques. This structured framework is essential for the development of detection tools that can address various types of fraud encountered in real-world situations.

2.2.1 ASVspoof 2021: Accelerating Progress in Spoofed and Deepfake Speech Detection.

[5] presents significant advancements in the field of audio spoofing and deepfake speech detection. ASVspoof challenges have been pivotal in driving the research and development of countermeasures against spoofed and deepfake audio attacks. The 2021 edition of this challenge focused on accelerating progress in the detection of such malicious audio manipulations, providing a comprehensive benchmark for evaluating various detection systems.

The ASVspoof 2021 challenge introduced new datasets and protocols designed to test the robustness and generalization of spoofing detection systems. One of the primary goals of this challenge was to bridge the gap between research and real-world application by providing more diverse and realistic spoofing scenarios. The datasets included a wide range of spoofing attacks, from traditional speech synthesis and voice conversion techniques to more advanced deepfake methods that utilize neural networks to generate highly convincing fake audio.

A key contribution of the ASVspoof 2021 challenge was the incorporation of both logical access (LA) and physical access (PA) scenarios. The LA scenario centers on attacks where the manipulated audio is directly fed into the automatic speaker verification (ASV) system, typically through digital channels. Conversely, the PA scenario involves replay attacks, wherein the altered audio is played in a physical setting and captured by the ASV system's microphone. This dual-pronged strategy ensures that detection systems are scrutinized across a broader spectrum of attack vectors, mirroring the intricacies encountered in real-world settings.

The challenge also placed a strong emphasis on deepfake detection, recognizing the growing threat posed by generative adversarial networks (GANs) and other neural network-based techniques capable of producing highly realistic fake audio. By incorporating deepfake speech into the evaluation, ASVspoof 2021 aimed to push

the boundaries of existing detection methodologies and encourage the development of more sophisticated and resilient countermeasures.

In terms of evaluation metrics, the ASVspoof 2021 challenge adopted the tandem detection cost function (t-DCF), which combines the detection error trade-off (DET) curve with the cost of spoofing attacks on ASV systems. This metric provides a more holistic assessment of the performance of spoofing detection systems by considering both the false alarm rate and the impact of spoofing on the overall security of ASV systems. By using t-DCF, the challenge promotes the development of detection systems that not only achieve high accuracy but also effectively mitigate the risks associated with spoofed and deepfake audio.

The results of the ASVspoof 2021 challenge highlighted several promising approaches to spoofing and deepfake detection. Notably, systems that leveraged advanced machine learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), demonstrated superior performance. These models were able to capture intricate patterns and temporal dependencies in audio signals, making them highly effective at distinguishing between genuine and spoofed speech. Additionally, the integration of data augmentation and adversarial training strategies helped improve the robustness of detection systems against a wide range of attacks.

In conclusion, the ASVspoof 2021 challenge has made significant strides in advancing the field of audio spoofing and deepfake speech detection. By providing a comprehensive and realistic benchmark, it has facilitated the development and evaluation of more effective detection systems. The inclusion of diverse attack scenarios and the emphasis on deepfake detection have pushed the boundaries of existing methodologies, encouraging the research community to develop more robust and resilient countermeasures. The challenge's impact extends beyond academic research, offering valuable insights and tools for enhancing the security of ASV systems in real-world applications.

2.2.2 Deepfake Audio Detection via MFCC Features Using Machine Learning

Presents a novel approach to identifying deepfake audio using Mel-Frequency Cepstral Coefficients (MFCCs) combined with machine learning techniques. Deepfake audio, which involves generating or manipulating audio recordings using advanced algorithms such as neural networks, poses significant threats to security, privacy, and authenticity. This research aims to address these concerns by developing a robust detection framework.

Mel-Frequency Cepstral Coefficients (MFCCs)

MFCCs are a representation of the short-term power spectrum of sound and have been extensively used in various speech and audio processing applications. They capture the perceptual aspects of audio signals, making them highly effective for distinguishing between different types of sounds, including speech and non-speech elements. In the context of deepfake audio detection, MFCCs serve as critical features that help identify subtle discrepancies introduced by generative models.

Methodology

The study utilizes a dataset comprising both genuine and deepfake audio samples. The MFCCs are extracted from these audio files to serve as input features for the machine learning models. The process involves the following steps:

1. **Preprocessing:** Audio samples are normalized and segmented into short frames. Each frame undergoes a Fourier Transform to convert the time-domain signals into the frequency domain.

2. **MFCC Extraction:** MFCCs are computed for each frame, capturing the spectral properties of the audio. This step involves filtering the frequency domain representation using a set of triangular bandpass filters that mimic the human ear's response.

3. **Feature Aggregation:** The MFCCs from all frames are aggregated to form a feature vector representing the entire audio sample. This vector serves as the input to the machine learning models.

Machine Learning Models

Several machine learning models are evaluated for their effectiveness in detecting deepfake audio, including:

- **Support Vector Machines (SVM)**
- **Random Forest (RF)**
- **Convolutional Neural Networks (CNN)**
- **Long Short-Term Memory (LSTM)**

Results

The experiments demonstrate that MFCCs are highly effective features for deepfake audio detection. Among the machine learning models tested, CNNs and LSTMs show superior performance due to their ability to capture complex patterns and temporal dependencies in the audio data. The study reports high accuracy rates, with CNNs achieving the best performance, followed closely by LSTMs. The use of MFCCs as input features significantly enhances the detection capabilities compared to traditional audio features.

Discussion

The research highlights the importance of feature selection in deepfake audio detection. MFCCs, due to their perceptual relevance, provide a robust basis for distinguishing between genuine and manipulated audio. The success of CNNs and LSTMs underscores the need for models that can capture both spatial and temporal characteristics of audio signals. However, the study also notes the computational complexity associated with deep learning models, suggesting the need for optimization techniques to ensure practical deployment in real-time applications.

Future Work

The paper outlines several directions for future research:

- Enhanced Feature Extraction: Exploring other feature extraction techniques, such as Mel-spectrograms and chroma features, to complement MFCCs and improve detection accuracy.
- Hybrid Models: Combining multiple machine learning models to leverage their strengths and mitigate their weaknesses.

- Real-World Scenarios: Testing the detection framework in diverse real-world scenarios, including noisy environments and different languages.
- Adversarial Robustness: Developing methods to enhance the robustness of detection systems against adversarial attacks that attempt to fool the detection algorithms.

Conclusion

The study "Deepfake Audio Detection via MFCC Features Using Machine Learning" provides a comprehensive approach to addressing the challenges of deepfake audio detection. By leveraging MFCCs and advanced machine learning models, the research offers a promising solution to detect and mitigate the threats posed by deepfake audio, ensuring greater security and authenticity in audio communications.

2.2.3 Siamese Network-Based MultiModal DeepFake Detection.

[**20**] introduces a novel approach to addressing the challenge of detecting deepfake content across multiple modalities. The paper leverages Siamese networks, a type of neural network architecture known for its ability to learn robust representations by comparing pairs of inputs, to enhance the detection accuracy of deepfake media.

The study is motivated by the increasing sophistication of deepfake techniques, which pose significant threats to various applications, including security, privacy, and misinformation detection. By integrating information from multiple modalities such as audio, video, and textual metadata, the proposed approach aims to improve the robustness and reliability of deepfake detection systems.

One of the key contributions of this research lies in its focus on multimodal fusion techniques. By combining features extracted from different modalities using Siamese networks, the model effectively learns to distinguish between genuine and manipulated content across diverse types of media. This approach not only enhances the detection accuracy but also strengthens the resilience of the system against adversarial attacks and emerging deepfake generation methods.

The evaluation methodology employed in the paper includes benchmarking against standard deepfake datasets and comparing performance metrics with state-of-the-art detection systems. The results demonstrate that the Siamese network-based approach achieves competitive performance in terms of accuracy and robustness, particularly when confronted with complex multimodal deepfake scenarios.

Furthermore, the paper discusses the implications of its findings for both research and practical applications. By advancing the state-of-the-art in multimodal deepfake detection, the study contributes to mitigating the risks associated with increasingly sophisticated digital manipulation techniques. It underscores the importance of integrating multimodal information to enhance the reliability of detection systems in real-world settings, where multimedia content is diverse and rapidly evolving.

In conclusion, "Siamese Network-Based MultiModal DeepFake Detection" presents a significant advancement in the field, leveraging Siamese networks and multimodal fusion techniques to detect deepfake content across various media types. The study not only improves upon existing methodologies but also sets a foundation for future research in enhancing the security and authenticity of multimedia content in digital environments.

CH 3 : Research Methodology

3.1 Overview

This chapter describes the research methods used to develop and evaluate the spoofed audio detection and speaker identification system. The methodology consists of several key stages, including dataset collection and preparation, feature extraction, model training and evaluation, and system integration.

First, diverse datasets containing both genuine and spoofed audio samples are collected and curated to ensure comprehensive coverage of various manipulation techniques, particularly those categorized by the ASVspoof Challenge (A0 to A19). This includes gathering audio data from public sources, proprietary databases, and simulated attacks.

Next, advanced feature extraction techniques are employed to capture the distinctive characteristics of audio signals. This step involves extracting features such as Mel-Frequency Cepstral Coefficients (MFCCs) and Mel-spectrogram features, which are crucial for differentiating between genuine and spoofed audio.

Following feature extraction, ten types of deep learning models are developed and tested: Long Short-Term Memory (LSTM) model, multi head transformer, WaveNet, Resnet, u-net, xceptionNet, VGGish, wave-u-net, GRU, and wave2vec. These models are chosen for their ability to handle sequential data and generate high-quality audio representations, respectively.

The models are trained and evaluated using the collected datasets, with performance metrics such as accuracy and EER used to assess their effectiveness. Special attention is given to the system's ability to generalize across different types of attacks and its resilience to adversarial examples.

Finally, the chosen models are integrated into a functional pipeline with a user-friendly interface. This system is designed to be easily deployable within banking authentication frameworks and adaptable to other voice-driven applications. The methodology ensures a systematic approach to developing a robust and reliable audio authentication system capable of addressing current and emerging threats.

Steps for Computing MFCCs

Step 1: Pre-Emphasis

The first step is to apply a pre-emphasis filter to amplify the high frequencies. This can be achieved with a simple high-pass filter:

$$y[n] = x[n] - \alpha x[n-1]$$

where $x[n]$ is the input signal, $y[n]$ is the output signal, and α is typically around 0.97.

Step 2: Framing

The signal is divided into short frames of equal length. If the signal x has N samples and the frame length is M, with an overlap of O, the frames are defined as:

$$x_i = x[i \times (M - O) : i \times (M - O) + M]$$

where i is the frame index.

Step 3: Windowing

Each frame is multiplied by a window function, typically a Hamming window:

$$x_i[n] = x_i[n] \cdot w[n]$$

where

$$w[n] = 0.54 - 0.46 \cos\left(\frac{2\pi n}{M-1}\right)$$

.

Step 4: Fast Fourier Transform (FFT)

Compute the FFT of each windowed frame to convert it from the time domain to the frequency domain:

$$X_i[k] = \sum_{n=0}^{M-1} x_i[n] e^{-j\frac{2\pi kn}{M}}$$

where $k = 0, 1, \ldots, M-1$.

Step 5: Power Spectrum

Calculate the power spectrum of the frames:

$$P_i[k] = \frac{|X_i[k]|^2}{M}$$

.

Step 6: Mel Filter Bank

Apply a set of Mel filter banks to the power spectra. The filters are triangular and spaced according to the Mel scale. The Mel scale $m(f)$ is given by:

$$m(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

The filter bank output S_m for the m-th filter is:

$$S_m = \sum_{k=0}^{K-1} P_i[k] H_m[k]$$

where $H_m[k]$ is the response of the m-th Mel filter at frequency bin k.

Step 7: Logarithm

Take the logarithm of each of the Mel filter bank energies:

$$L_m = \log(S_m)$$

.

Step 8: Discrete Cosine Transform (DCT)

Apply the Discrete Cosine Transform to the log filter bank energies to get the MFCCs:

$$C_n = \sum_{m=0}^{M-1} L_m \cos\left[\frac{\pi n(2m+1)}{2M}\right]$$

where C_n represents the n-th MFCC. Usually, the first few coefficients are taken to form the feature vector.

Summary Formula

To summarize the process in a simple formula:

$$\text{MFCCs} = \text{DCT}\left(\log\left(\text{MelFilterBank}\left(\text{PowerSpectrum}\left(\text{FFT}\left(\text{Window}\left(\text{Frames}\left(\text{Pre-Emphasis}\left(x\right)\right)\right)\right)\right)\right)\right)\right)$$

Additionally, the MFCCs and Mel-spectrogram features are fused using Principal Component Analysis (PCA) for dimensionality reduction and enhanced feature representation. This sequence of steps and transformations efficiently extracts features from an audio signal, capturing the spectral characteristics in a form that is useful for various audio processing tasks, such as speech recognition and audio classification.

3.2 Dataset

3.2.1 ASVspoof 2019 LA (Logical Access)

Overview: The ASVspoof 2019 Logical Access (LA) dataset is a widely used benchmark for evaluating audio spoofing detection systems.

Purpose: It focuses on detecting synthetic speech, which includes both text-to-speech (TTS) and voice conversion (VC) attacks.

Content: The dataset contains genuine and spoofed speech samples generated using various state-of-the-art TTS and VC methods.

Structure: It is split into training, development, and evaluation sets, facilitating the development and benchmarking of detection systems.

Significance: Provides a standardized and challenging set of spoofing attacks, promoting the development of robust detection algorithms.

3.2.2 WaveFake

Overview: WaveFake is a dataset specifically designed to evaluate the robustness of spoofing detection systems against audio deepfakes.

Purpose: It aims to cover a diverse range of synthetic speech generated by different deep learning models.

Content: The dataset includes audio samples generated using various generative models, such as WaveNet, WaveRNN, and other GAN-based models.

Structure: It includes both genuine speech and synthetic samples, with clear labels for each.

Significance: Helps in understanding the effectiveness of detection systems against a wide array of deepfake generation techniques.

3.2.3 DEEP-VOICE: DeepFake Voice Recognition

Overview: DEEP-VOICE is another dataset curated to tackle the challenges of detecting deepfake voices.

Purpose: It provides a diverse set of deepfake audio samples to enhance the detection capabilities of systems against advanced spoofing techniques.

Content: The dataset consists of genuine and deepfake audio samples, with a focus on high-quality, realistic voice synthesis.

Structure: It is organized into different subsets for training and testing, ensuring comprehensive evaluation.

Significance: Contributes to the development of more sophisticated detection methods by offering complex and realistic deepfake examples.

3.2.4 Key Points to Highlight

Variety and Diversity: Using multiple datasets ensures that the detection system is robust against a wide range of spoofing techniques, that is achieved through sampling and balancing the data as for ASVspoof, the data had 91% real against only 9% fake utterances in the data however the data had a good distribution for transmission channels and attacking methods, as for wavefake that had 10 different sophisticated generating method. As the datasets distribution shown in Figure 4, and to sum up the data was gathered 2000 sample from each method from wavefake, 1500 from each 6 transmision and codec method in asvspoof across the 20 attack type, and 1000 from deep-voice for a total of 40,000 sample.

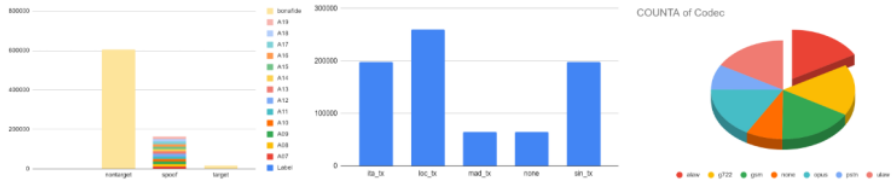


Figure 4: Datasets distribution

Benchmarking: These datasets are well-recognized benchmarks in the industry and the research community, providing a solid foundation for evaluating and comparing detection algorithms.

Realism and Complexity: The datasets include high-quality and realistic spoofing examples, challenging the detection system to be more effective.

3.3 System architecture

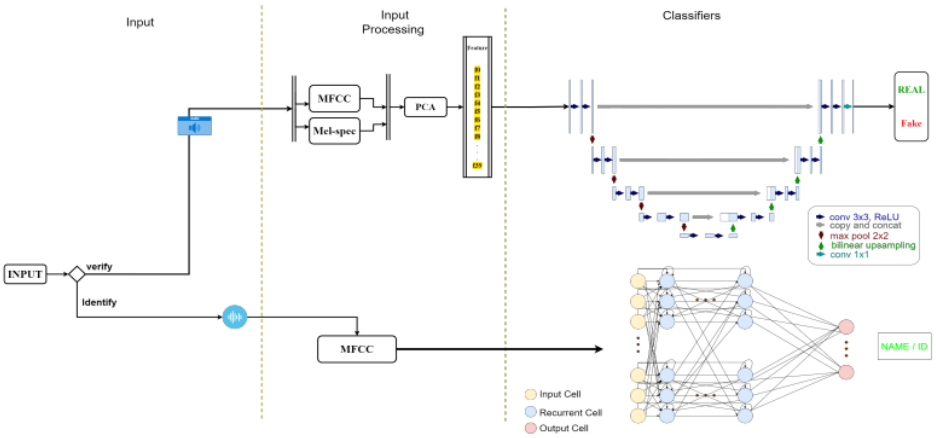


Figure 5: System architecture

Ten different deep models were employed to develop and test the audio spoofing detection models: Long Short-Term Memory (LSTM) model, multi head transformer, WaveNet, Resnet, u-net, ceptionNet, VGGish, wave-u-net, GRU, and wave2vec. Each architecture was trained and evaluated using both the *ASVspoof 2019 LA* dataset and a refined dataset combining samples from all three datasets.

3.3.1 Long Short-Term Memory (LSTM)

LSTM networks are well-suited for sequence prediction problems and have been widely used in speech processing tasks. The LSTM model was trained and tested on both datasets to evaluate its performance in detecting audio spoofing.

LSTM models are particularly effective in handling the temporal dependencies in audio data, which is critical for distinguishing between genuine and spoofed audio samples. Their architecture, which includes memory cells and gating mechanisms, allows them to maintain and update information over long sequences, thereby capturing the temporal patterns and nuances essential for audio spoofing detection. In this study, the LSTM model was fine-tuned to optimize its performance on the audio datasets. Various hyperparameters, including the number of LSTM layers, the number of units per layer, and the learning rate, were experimented with to achieve the best possible performance.

During the training phase, the LSTM model was exposed to a diverse set of audio samples from the *ASVspoof 2019 LA* dataset and the refined dataset. This exposure helped the model learn to differentiate between legitimate and spoofed audio based on features extracted from the sequential data. The evaluation metrics included accuracy, precision, recall, and F1 score, which provided a comprehensive assessment of the model's performance.

The results indicated that the LSTM model was effective in identifying audio spoofing, although its performance varied depending on the complexity and variability of the audio samples in the datasets. Despite its effectiveness, the LSTM model's training time and computational requirements were higher compared to

some other models, highlighting the trade-off between accuracy and efficiency in using LSTM networks for this task.

3.3.2 Transformer

Transformer networks have gained popularity for their effectiveness in various natural language processing tasks. Their self-attention mechanism enables the model to focus on different parts of the input sequence, making them suitable for audio spoofing detection.

The self-attention mechanism of transformers allows the model to weigh the importance of different parts of the input sequence, capturing both local and global dependencies within the audio data. This characteristic is particularly advantageous for audio spoofing detection, where subtle differences in the temporal structure of audio signals can indicate spoofing.

In this study, the transformer model was trained on the *ASVspoof 2019 LA* dataset and the refined dataset, utilizing its ability to process sequences in parallel and handle long-range dependencies efficiently. The model's architecture includes multiple attention heads, which enable it to learn different aspects of the audio data simultaneously, improving its ability to detect nuanced patterns that distinguish spoofed audio from genuine samples.

Hyperparameter tuning for the transformer model involved adjusting the number of layers, the size of the hidden states, the number of attention heads, and the learning rate. These parameters were optimized to enhance the model's performance in audio spoofing detection tasks.

Evaluation of the transformer model demonstrated its robustness in handling diverse audio samples and its capability to achieve high accuracy in detecting spoofed audio. The transformer model's ability to parallelize computations resulted in faster training times compared to sequential models like LSTM, making it a more efficient choice for large-scale audio spoofing detection tasks. However, the model's performance was sensitive to the quality and variability of the training data, underscoring the importance of a well-curated dataset for achieving optimal results.

3.3.3 WaveNet

WaveNet models, based on convolutional neural networks, are designed to generate raw audio waveforms and have shown excellent performance in speech synthesis and audio processing tasks. Their architecture leverages dilated causal convolutions, which allow the model to capture long-range temporal dependencies in the audio signal without the need for recurrent connections. This makes WaveNet particularly effective at modeling complex patterns in audio data.

In the context of audio spoofing detection, WaveNet's ability to generate detailed and high-quality audio waveforms is utilized to distinguish between genuine and spoofed audio samples. The model's deep convolutional layers and hierarchical structure enable it to learn intricate features of the audio signal, thereby improving its accuracy in detecting subtle anomalies that indicate spoofing. WaveNet was trained and evaluated on both the *ASVspoof 2019 LA* dataset and the refined dataset, demonstrating its robustness and effectiveness in this task.

3.3.4 Resnet

The ResNet model (Residual Network) is known for its deep structure, which reduces the problem of gradient disappearance through skip connections. These skip connections allow the model to carry forward information across layers without degradation, enabling the training of much deeper networks. This capability makes ResNet particularly effective for complex tasks requiring deep feature extraction.

In the context of audio spoofing detection, ResNet models are suitable because they can learn dynamic features from audio spectrograms. By leveraging its deep architecture, ResNet can capture intricate patterns and temporal dependencies within the audio data that are indicative of spoofing. The model was trained and tested on both the *ASVspoof 2019 LA* dataset and the refined dataset, showcasing its ability to generalize well across different types of audio spoofing attacks. The effectiveness of ResNet in this domain is attributed to its powerful feature learning capabilities, making it a robust choice for detecting subtle anomalies in audio signals.

3.3.5 U-Net

U-Net is a convolutional neural network originally designed for biomedical image segmentation. Its encoder-decoder structure allows it to capture both global and local features, making it effective for tasks requiring precise feature localization, such as audio spoofing detection. The encoder part of the U-Net model down-samples the input to extract high-level features, while the decoder up-samples these features to produce an output with the same spatial dimensions as the input.

In audio spoofing detection, U-Net can effectively learn and represent the complex temporal and spectral patterns present in audio signals. The model's ability to localize features precisely within the audio spectrogram makes it particularly well-suited for identifying subtle cues that differentiate spoofed audio from genuine audio. By leveraging both global context and fine details, U-Net provides a comprehensive approach to detecting audio spoofing attacks. The model was trained and evaluated on the *ASVspoof 2019 LA* dataset and the refined dataset, demonstrating its robustness and accuracy in various audio spoofing scenarios.

3.3.6 XceptionNet

XceptionNet is an extension of the Inception architecture, which replaces the standard Inception modules with depthwise separable convolutions. This modification allows the model to perform more efficient and precise convolutions, leading to better performance with fewer parameters. XceptionNet has been widely successful in image classification tasks due to its ability to capture intricate patterns in data.

For audio spoofing detection, XceptionNet's architecture can effectively handle the complex and high-dimensional nature of audio spectrograms. By utilizing depthwise separable convolutions, the model can focus on channel-wise correlations and spatial correlations separately, enhancing its ability to distinguish between genuine and spoofed audio. The model was trained on both the *ASVspoof 2019 LA* dataset and the combined dataset, where it demonstrated significant potential in accurately identifying audio spoofing attacks with its robust feature extraction capabilities.

3.3.7 VGGish

VGGish is a variant of the VGG network specifically adapted for audio classification tasks. It uses the same principles of deep convolutional layers to learn hierarchical feature representations but is tailored to handle audio data by transforming it into a Mel-spectrogram input. This process involves converting raw audio signals into a visual representation that the convolutional layers can process effectively.

In the context of audio spoofing detection, VGGish's architecture allows it to capture the essential temporal and spectral features necessary for distinguishing between authentic and spoofed audio. Despite its relatively large size of 1.12 gigabytes, VGGish has shown strong performance due to its deep and wide network structure, which enables it to learn from vast amounts of audio data. The model was trained on the *ASVspoof 2019 LA* dataset and the combined dataset, where it exhibited high accuracy and robustness in detecting spoofed audio signals.

3.3.8 Wave-U-Net

Wave-U-Net combines the concepts of WaveNet and U-Net, leveraging convolutional architectures to model raw audio waveforms. This model is particularly adept at capturing both temporal and spectral features of audio, providing a comprehensive approach to spoofing detection.

3.3.9 GRU

Gated Recurrent Unit (GRU) networks are similar to LSTMs but with a simplified structure that combines the forget and input gates into a single update gate. GRUs have been effective in sequence modeling tasks, offering a balance between model complexity and performance in audio spoofing detection.

3.3.10 Wave2Vec

Wave2Vec is a model designed for unsupervised representation learning of raw audio. By training on large amounts of unlabeled audio data, it learns useful features that can be fine-tuned for various downstream tasks, including audio spoofing detection. Wave2Vec's ability to leverage unlabeled data makes it a powerful tool for improving detection accuracy in diverse audio environments.

4.1 Experimental design

Each model was trained and tested using both the *ASVspoof 2019 LA* dataset and the refined dataset, which combines samples from all three datasets. The training involved optimizing the models using appropriate loss functions and performance metrics, such as accuracy. The testing phase evaluated the models' ability to generalize to unseen data, providing a comprehensive assessment of their performance.

Model	LSTM	Resnet	M-Transformer	wavenet
ASV	74.68	91.63	91.38	86.24
refined	92.67	93.64	92.29	92.99
ref/asv	89.70	94.78	95.63	94.01

Model	u-net	xceptionNet	VGGish	wave-u-net	GRU	wave2vec
ASV	82.34	82.39	93.81	67.36	-	92.49
refined	93.53	93.71	96.27	94.63	80.61	91.29
ref/asv	95.17	91.65	97.86	95.23	78.1	93.90

Table 1: Model Performance Comparison

The methodology adopted in this research involved using diverse and challenging datasets, experimenting with different model architectures, and rigorous training and testing procedures. This approach ensured the development of robust and effective audio spoofing detection models, capable of generalizing across various types of synthetic speech attacks.

4.2 Results & discussion

Based on the experimental results, the U-net model was chosen as the most suitable for our application. Although some other models demonstrated superior performance in specific metrics, they had notable drawbacks. For instance, the VGGish model exhibited high performance but its size was 1.12 gigabytes, which is impractical for our resource-constrained environment. On the other hand, the wave-U-net model achieved an impressive 99.58% accuracy on the testing set. However, this exceptionally high accuracy suggests a high variance, indicating that the model might be overfitting the training data.

In contrast, the U-net model provided a balanced trade-off between performance and practicality. While it may not have achieved the highest accuracy among the tested models, its size and generalization capability make it the most viable option for deployment. The U-net's architecture allows for efficient processing without compromising too much on accuracy, making it the optimal choice considering both performance and resource limitations.

5.1 Review

The development of a robust audio deep-fake detection system for banking applications has been a complex yet rewarding journey. This project aimed to address the critical need for secure and reliable speaker verification systems capable of identifying sophisticated audio spoofing attacks. Throughout the research, we explored various model architectures, including Long Short-Term Memory (LSTM) networks, Transformer models, and WaveNet models, and evaluated their performance using comprehensive datasets such as ASVspoof 2019 LA and WaveFake.

The U-net model emerged as the most suitable choice for our application, balancing performance with practical considerations such as model size. Although other models like VGGish demonstrated higher accuracy in certain metrics, their substantial size rendered them less practical for resource-constrained environments typical of banking systems. The wave-U-net, despite its high accuracy on testing, indicated a high variance, which could undermine its reliability in real-world scenarios.

The research methodology employed rigorous training and testing procedures, leveraging advanced feature extraction techniques and robust datasets. This systematic approach ensured the development of a resilient audio authentication system capable of generalizing across various manipulation methods and scenarios.

In conclusion, the findings of this research underscore the importance of integrating sophisticated machine learning models with practical considerations to enhance the security of voice-driven banking systems. Future work will focus on further refining these models and exploring additional features to anticipate and mitigate emerging threats in the field of audio deep-fakes.

5.2 Limitations

Despite the promising results and potential applications of the developed audio deep-fake detection system, several limitations were identified during the research and development process.

1 - Scalability of the Identification Model:

The current identification model is based on Long Short-Term Memory (LSTM) networks. While this model has demonstrated satisfactory performance on a small number of customers, its scalability is a concern. As the number of customers increases, the LSTM model may struggle to maintain its performance and efficiency. This indicates a need for a more scalable model that can handle larger datasets and more complex scenarios without compromising on accuracy or speed.

2 - Adding New Customers:

The existing system requires transfer learning and retraining of the model to incorporate new customers. This approach is neither efficient nor practical for a real-world application where the customer base is constantly changing and expanding. A more dynamic and seamless method for adding new customers is necessary to ensure the system's usability and effectiveness in a live environment.

These limitations highlight the need for further research and development to address the scalability issues of the LSTM model and to devise a more efficient mechanism for integrating new customers. Future work will focus on exploring al-

ternative models and methodologies that can overcome these challenges and enhance the system's overall performance and adaptability.

5.3 Future work

To address the limitations identified, several avenues for future work are proposed:
1 - Development of a Scalable Identification Model:
Future research should focus on developing or integrating a more scalable model that can handle a larger and more diverse customer base. Exploring models such as Transformer-based architectures or hybrid models that combine the strengths of LSTM with other techniques may provide a solution to the scalability issue.
2 - Efficient Customer Integration Mechanism:
A key area for improvement is the method for adding new customers to the system. Future work should aim to create an efficient and automated process for integrating new customers without the need for extensive retraining. Techniques such as few-shot learning or continuous learning models could be explored to facilitate this process.
3 - Enhanced Model Performance and Robustness:
Ongoing efforts should be made to improve the overall performance and robustness of the detection system. This includes refining the existing models to reduce variance and overfitting, as well as incorporating additional features and data sources to enhance the accuracy and reliability of the system.
4 - Real-World Testing and Validation:
Finally, future work should include extensive testing and validation of the system in real-world scenarios. This will help identify any additional challenges and ensure that the system is robust, reliable, and ready for deployment in practical applications.

1 Piotr Kawa, Marcin Plata, Piotr Syga. *SpecRNet: Towards Faster and More Accessible Audio DeepFake Detection*. October 2022. Available at: `https://arxiv.org/abs/your-arxiv-id`.

2 Zaynab Almutairi, Hebah Elgibreen. *A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions*. May 2022.

3 Run Wang, Felix Juefei-Xu, Yihao Huang, Qing Guo, Xiaofei Xie, Lei Ma, Yang Liu. *DeepSonar: Towards Effective and Robust Detection of AI-Synthesized Fake Voices*. August 2020.

4 Tuba Arif, Ali Javed, Mohammed Alhameed, Fathe Jeribi, Ali Tahir. *Voice Spoofing Countermeasure for Logical Access Attacks Detection*. December 2021.

5 J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, H. Delgado. *ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection.* In ASVspoof 2021 Workshop - Automatic Speaker Verification and Spoofing Countermeasures Challenge. September 2021.

6 Ameer Hamza, Abdul Rehman Javed, Farkhund Iqbal, Natalia Kryvinska, Ahmad S. Almadhor, Zunera Jalil, Rouba Borghol. *Deepfake Audio Detection via MFCC Features Using Machine Learning*. December 2022.

7 Dmitry Efanov, Pavel Aleksandrov, Nikolay Karapetyants. *The BiLSTM-based synthesized speech recognition.* Procedia Computer Science, Volume 213, 2022.

8 T. Liu, D. Yan, R. Wang, N. Yan, G. Chen. *Identification of Fake Stereo Audio Using SVM and CNN.* Information, Volume 12, Issue 7, 2021, pp. 263.

9 Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, Dinesh Manocha. *Emotions Don't Lie: An Audio-Visual Deepfake Detection Method Using Affective Cues*. August 2020.

10 Joel Frank, Lea Schönherr. *Wavefake: A data set to facilitate audio deepfake detection.* arXiv preprint arXiv:2111.02813, 2021. Available at: `https://arxiv.org/abs/2111.02813`.

11 D. Cozzolino, A. Pianese, M. Nießner, L. Verdoliva. *Audio-Visual Person-of-Interest DeepFake Detection.* IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 2023, pp. 943-952. doi: 10.1109/CVPRW59228.2023.00

12 Birdy654. *Deep Voice DeepFake Voice Recognition Dataset.* Available at: `https://www.kaggle.com/datasets/birdy654/deep-voice-deepfake-voice-recognition`.

13 J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, H. Delgado. (2021). ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. In ASVspoof 2021 Workshop - Automatic Speaker Verification and Spoofing Countermeasures Challenge (September).

14 Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, Dinesh Manocha. (August 2020). Emotions Don't Lie: An Audio-Visual Deepfake Detection Method Using Affective Cues.

15 Joel Frank, Lea Schönherr. (2021). Wavefake: A data set to facilitate audio deepfake detection. arXiv preprint arXiv:2111.02813.

16 D. Cozzolino, A. Pianese, M. Nießner, L. Verdoliva. (June 2023). Audio-Visual Person-of-Interest DeepFake Detection. In IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (pp. 943-952). doi:

10.1109/CVPRW[illegible]228.2023.00101.

17 Birdy654. Deep Voice DeepFake Voice Recognition Dataset. https://www.kaggle.com/datasets voice-deepfake-voice-recognition.

18 Dmitry Efanov, Pavel Aleksandrov, Nikolay Karapetyants. (2022). The BiLSTM-based synthesized speech recognition. Procedia Computer Science, Volume 213.

19 Tianyun Liu, Diqun Yan, Rangding Wang, Nan Yan, Gang Chen. (2021). Identification of Fake Stereo Audio Using SVM and CNN. Information, Volume 12, Issue 7, pp. 263.

20 Raju Nekadi. (2020). Siamese Network-Based MultiModal DeepFake Detection.

21 Xiaoyang Tan, Bill Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions.

22 A. Pianese, D. Cozzolino, G. Poggi, L. Verdoliva. (2022). Deepfake audio detection by speaker verification.

23 R. L. M. A. P. C. Wijethunga, D. M. K. Matheesha, A. A. Noman, K. H. V. T. A. De Silva, M. Tissera, L. Rupasinghe. (2020). ADD 2022: THE FIRST AUDIO DEEP SYNTHESIS DETECTION CHALLENGE.

24 M. Hafizur Rahman, M. Graciarena, D. Castan, C. Cobo-Kroenke, M. McLaren, A. Lawson. (2022). Detecting Synthetic Speech Manipulation in Real Audio Recordings.

25 J. Yi et al. (2022). ADD 2022: THE FIRST AUDIO DEEP SYNTHESIS DETECTION CHALLENGE.

gradDoc

ORIGINALITY REPORT

19%	16%	13%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	export.arxiv.org Internet Source	2%
2	arxiv.org Internet Source	1%
3	www.mdpi.com Internet Source	1%
4	Submitted to Liverpool John Moores University Student Paper	1%
5	pure.tue.nl Internet Source	1%
6	www.ijirset.com Internet Source	1%
7	ceur-ws.org Internet Source	1%
8	ijisae.org Internet Source	1%
9	dblp.uni-trier.de Internet Source	1%

10	www.ftc.gov Internet Source	<1 %
11	ebin.pub Internet Source	<1 %
12	digibug.ugr.es Internet Source	<1 %
13	Xiaojie Mu, Cheol-Hong Min. "MFCC as Features for Speaker Classification using Machine Learning", 2023 IEEE World AI IoT Congress (AIIoT), 2023 Publication	<1 %
14	Submitted to University of Newcastle Student Paper	<1 %
15	deepai.org Internet Source	<1 %
16	Submitted to Caucasus University Student Paper	<1 %
17	Khaing Zar Mon, Kasorn Galajit, Candy Olivia Mawalim, Jessada Karnjana, Tsuyoshi Isshiki, Pakinee Aimmanee. "Spoof Detection using Voice Contribution on LFCC features and ResNet-34", 2023 18th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), 2023 Publication	<1 %
	Submitted to University of Surrey	

18	Student Paper	<1%
19	link.springer.com Internet Source	<1%
20	"Soft Computing and Signal Processing", Springer Science and Business Media LLC, 2019 Publication	<1%
21	Submitted to Higher Education Commission Pakistan Student Paper	<1%
22	par.nsf.gov Internet Source	<1%
23	Wenyuan Yang, Xiaoyu Zhou, Zhikai Chen, Bofei Guo, Zhongjie Ba, Zhihua Xia, Xiaochun Cao, Kui Ren. "AVoiD-DF: Audio-Visual Joint Learning for Detecting Deepfake", IEEE Transactions on Information Forensics and Security, 2023 Publication	<1%
24	assets.researchsquare.com Internet Source	<1%
25	Submitted to University of Pittsburgh Student Paper	<1%
26	Kuldeep Chouhan, Abhishek Singh, Anurag Shrivastava, Shweta Agrawal, Brahma Datta	<1%

Shukla, Pragya Singh Tomar. "Structural Support Vector Machine for Speech Recognition Classification with CNN Approach", 2021 9th International Conference on Cyber and IT Service Management (CITSM), 2021
Publication

27	dokumen.pub Internet Source	<1 %
28	dspace.univ-ouargla.dz Internet Source	<1 %
29	Submitted to Manchester Metropolitan University Student Paper	<1 %
30	www.isca-speech.org Internet Source	<1 %
31	Abdrabuh, Ehab Alsayed Albadawy. "AI-Synthesized Speech: Generation and Detection", State University of New York at Albany, 2022 Publication	<1 %
32	Piotr Kawa, Marcin Plata, Piotr Syga. "SpecRNet: Towards Faster and More Accessible Audio DeepFake Detection", 2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), 2022 Publication	<1 %

33	scholar.archive.org Internet Source	<1%
34	Ton Duc Thang University Publication	<1%
35	Submitted to University of Auckland Student Paper	<1%
36	Albérick Euraste Djiré, Aminata Sabané, Abdoul-Kader Kabore, Rodrique Kafando, Tégawendé F. Bissyandé. "Evaluating Acoustic Parameters for DeepFake Audio Identification", 2023 IEEE Afro-Mediterranean Conference on Artificial Intelligence (AMCAI), 2023 Publication	<1%
37	Submitted to University of Sussex Student Paper	<1%
38	Yue Zhang, Zimo Zhou, Ying Deng, Daiwei Pan, Jesse Van Griensven Thé, Simon X. Yang, Bahram Gharabaghi. "Daily Streamflow Forecasting Using Networks of Real-Time Monitoring Stations and Hybrid Machine Learning Methods", Water, 2024 Publication	<1%
39	vigir.missouri.edu Internet Source	<1%
	Submitted to Heidelberg University	

40	Student Paper	<1%
41	Yeqing Ren, Haipeng Peng, Lixiang Li, Xiaopeng Xue, Yang Lan, Yixian Yang. "Generalized Voice Spoofing Detection via Integral Knowledge Amalgamation", IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023 Publication	<1%
42	"Data Science and Applications", Springer Science and Business Media LLC, 2024 Publication	<1%
43	5dok.net Internet Source	<1%
44	Submitted to University of Sheffield Student Paper	<1%
45	Submitted to University of Wolverhampton Student Paper	<1%
46	Submitted to universititeknologimara Student Paper	<1%
47	Alyammahi, Abdulla Jumah. "Exploring the Role of Co-Production in Enhancing Happiness and Well-Being in the UAE.", The British University in Dubai, 2023 Publication	<1%

48	Ousama A Shaaban, Remzi Yildirim, Abubaker Alguttar. "Audio Deepfake Approaches", IEEE Access, 2023 Publication	<1 %
49	Sagar Nailwal, Saksham Singhal, Nongmeikapam Thoiba Singh, Arbaz Raza. "Deepfake Detection: A Multi-Algorithmic and Multi-Modal Approach for Robust Detection and Analysis", 2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE), 2023 Publication	<1 %
50	acikbilim.yok.gov.tr Internet Source	<1 %
51	ir.knust.edu.gh Internet Source	<1 %
52	Submitted to South Bank University Student Paper	<1 %
53	ruor.uottawa.ca Internet Source	<1 %
54	Il-Youp Kwak, Sunmook Choi, Jonghoon Yang, Yerin Lee, Soyul Han, Seungsang Oh. "Low-quality Fake Audio Detection through Frequency Feature Masking", Proceedings of	<1 %

	the 1st International Workshop on Deepfake Detection for Audio Multimedia, 2022 Publication	
55	Tang, Xuting. "Trust in the AI-R: Accuracy, Interpretability, Resilience and Fairness", Stevens Institute of Technology, 2023 Publication	<1 %
56	Yao, Jiahao. "Reinforcement Learning and Variational Quantum Algorithms", University of California, Berkeley, 2024 Publication	<1 %
57	dspace02.jaist.ac.jp Internet Source	<1 %
58	ebuah.uah.es Internet Source	<1 %
59	i-rep.emu.edu.tr:8080 Internet Source	<1 %
60	ijettjournal.org Internet Source	<1 %
61	encyclopedia.pub Internet Source	<1 %
62	medium.com Internet Source	<1 %
63	sibgrapi.sid.inpe.br Internet Source	<1 %

64	"Medical Image Computing and Computer Assisted Intervention – MICCAI 2018", Springer Nature America, Inc, 2018 Publication	<1 %
65	Ameer Hamza, Abdul Rehman Javed, Farkhud Iqbal, Natalia Kryvinska, Ahmad S. Almadhor, Zunera Jalil, Rouba Borghol. "Deepfake Audio Detection via MFCC features using Machine Learning", IEEE Access, 2022 Publication	<1 %
66	Guoyuan Lin, Weiqi Luo, Da Luo, Jiwu Huang. "One-Class Neural Network with Directed Statistics Pooling for Spoofing Speech Detection", IEEE Transactions on Information Forensics and Security, 2024 Publication	<1 %
67	asmp-eurasipjournals.springeropen.com Internet Source	<1 %
68	ouci.dntb.gov.ua Internet Source	<1 %
69	syndelltech.com Internet Source	<1 %
70	www.scilit.net Internet Source	<1 %
71	"Computer Vision and Image Processing", Springer Science and Business Media LLC,	<1 %

	2021 Publication	
72	Submitted to University of Greenwich Student Paper	<1 %
73	Zhongjie Ba, Qing Wen, Peng Cheng, Yuwei Wang, Feng Lin, Li Lu, Zhenguang Liu. "Transferring Audio Deepfake Detection Capability across Languages", Proceedings of the ACM Web Conference 2023, 2023 Publication	<1 %
74	Zrar Kh. Abdul, Abdulbasit K. Al-Talabani. "Mel Frequency Cepstral Coefficient and its Applications: A Review", IEEE Access, 2022 Publication	<1 %
75	assets-eu.researchsquare.com Internet Source	<1 %
76	erepo.uef.fi Internet Source	<1 %
77	fastercapital.com Internet Source	<1 %
78	reports-archive.adm.cs.cmu.edu Internet Source	<1 %
79	umpir.ump.edu.my Internet Source	<1 %
80	vdoc.pub Internet Source	<1 %

81	web.archive.org Internet Source	<1%
82	www.ijisrt.com Internet Source	<1%
83	www.nowpublishers.com Internet Source	<1%
84	www.publishing.globalcsrc.org Internet Source	<1%
85	Burak Yüksel. "Comparative Analysis of LSTM Model in Predicting ETF Stock Prices for Different Sectors", Marmara Universitesi (Turkey), 2024 Publication	<1%
86	João Phillipe Cardenuto, Jing Yang, Rafael Padilha, Renjie Wan et al. "The Age of Synthetic Realities: Challenges and Opportunities", APSIPA Transactions on Signal and Information Processing, 2023 Publication	<1%
87	Xin Wang, Junichi Yamagishi. "Chapter 8 A Practical Guide toLogical Access Voice Presentation Attack Detection", Springer Science and Business Media LLC, 2022 Publication	<1%
88	"Mining Intelligence and Knowledge Exploration", Springer Science and Business	<1%

Media LLC, 2018
Publication

89 Massimo Leone. "The Spiral of Digital Falsehood in Deepfakes", International Journal for the Semiotics of Law - Revue internationale de Sémiotique juridique, 2023 <1%
Publication

90 Rami Mubarak, Tariq Alsboui, Omar Alshaikh, Isa Inuwa-Dute, Saad Khan, Simon Parkinson. "A Survey on the Detection and Impacts of Deepfakes in Visual, Audio, and Textual Formats", IEEE Access, 2023 <1%
Publication

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		