Journal of Information Systems Engineering and Management
2025, 10(44s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Optimized Voice Spoofing Detection Using One Class Learning to Combat Identity Theft and Fraud

[1]Serilda Victoria I, [2]Vibha Shree S, [3]Dr. G. Maria Kalavathy
[1]Student Information Technology Easwari Engineering College Chennai, India
sydneeserilda07@gmail.com
[2]Student Information Technology Easwari Engineering College Chennai, India
vibhashreesrinivasan@gmail.com
[3]Professor Information Technology Easwari Engineering College Chennai, India
maria_kalavathy@yahoo.co.in

ARTICLE INFO

Received: 18 Dec 2024

Revised: 10 Feb 2025

Accepted: 28 Feb 2025

ABSTRACT

Voice recognition is essential to secure authentication systems, with increasing digitization. But the rise of advanced spoofing attacks emphasizes significant flaws in modern technologies, which frequently fall short in the midst of emerging threats. By employing a One-Class Learning technique and focusing only on genuine voice samples to detect deviations indicative of spoofing, this work addresses these vulnerabilities. To increase the system's adaptability, the methodology combines autoencoders for pattern recognition with data augmentation techniques like pitch, speed fluctuations. High detection accuracy can be observed in the results, regardless of various difficult circumstances. The architecture provides a scalable solution by enabling deployment in resource-constrained devices.

Keywords: Voice Spoofing Detection, One Class Learning, ASV systems, Spectral Feature Extraction, Cross Entropy Loss, Doc-SoftMax Function.

INTRODUCTION

In the digital age, Voice recognition has emerged as a key authentication method , used in sectors like banking, healthcare, and government. However, the rise of AI-powered voice cloning and deep fake technologies has significantly increased the risk of voice spoofing attacks, allowing attackers to impersonate users and circumvent security measures. Because of their limited generalization to unidentified spoofing techniques and dependence on manually constructed features, traditional automatic speaker verification (ASV) systems are susceptible to these attacks. This paper offers a ResNet-based One-Class Learning system meant to identify spoofing attempts using just real voice samples for training, so removing the requirement for labeled spoofed datasets to handle these issues. The following are the main contributions of this work : Our model reduces the need for large labeled spoof datasets using One Class Learning by detecting spoof audio as anomalies, in contrast to traditional binary classifiers. ResNet-Based Feature Extraction which is suitable for low-resource settings, this lightweight student model is based on ResNet and is optimized for high-speed inference. Multi-Feature Spectral Analysis to improve detection accuracy under various acoustic conditions, we combine sophisticated Mel-Frequency Cepstral Coefficients (MFCC), Linear Frequency Cepstral Coefficients (LFCC), Bark Frequency Cepstral Coefficients (BFCC), Power Quality Cepstral Coefficients (PQCC) . The rest of the paper is organized as follows: Section II provides a comprehensive literature review of existing voice spoofing detection techniques. Section III details the proposed methodology, covering data preprocessing, feature extraction, model training, and evaluation. Section IV presents experimental results, followed by a comparative analysis. Finally, Section V discusses the limitations and future research directions.

OBJECTIVES

The main objective of this work is to develop a robust voice spoofing detection system that improves the security of voice authentication applications. By employing a one-class learning approach, the system will utilize only genuine voice samples during training to detect and flag spoofed voices as anomalies. This approach eliminates the need for a large dataset of spoofed voices, offering a scalable, efficient, and adaptable solution for detecting unauthorized access attempts. The system is designed to enhance protection against evolving spoofing techniques in real-world environments, providing a secure and reliable authentication method.

In order to greatly improve security in voice authentication applications, this work aims to design a voice spoofing detection system. By using only real voice samples for training and a one-class learning approach, the system does not require a large amount of spoofed voice data. In order to identify and flag spoof voices as anomalies, the system learns the patterns of real voice inputs. This approach guarantees scalability and adaptability in a variety of security contexts while also expediting the training process. Strong defense against increasingly complex voice spoofing techniques is also provided by its dependable and effective mechanism for detecting unauthorized access attempts.

RELATED WORK

Voice spoofing detection has drawn more attention due to the quick development of voice converter and synthetic speech creation methods. The section categorizes these into three categories: deep learning models, contemporary one-class learning frameworks, and conventional supervised methods. Conventional supervised models that use hand-crafted features rely on binary classification. The CQCC-GMM system[1] introduced by Todisco et al. served as a benchmark in the ASVspoof 2017 challenge. Despite it working well under controlled conditions, it was not resilient enough against unseen spoofing techniques. Similarly, linear models and Constant-Q Cepstral Coefficients (CQCCs)[2] showed considerable performance in ASVspoof 2019, but their reliance on labeled spoof samples makes them infeasible in real-world zero-day attack scenarios.

For spoofing detection, recent advancements have introduced CNNs, ResNets, and recurrent architectures. Lavrentyeva et al. introduced a Light CNN model[3] trained on the ASVspoof 2019 dataset. It exhibited high performance but required extensive labeled spoof data. Xia et al. improved channel attention[4] in spectrogram inputs by utilizing, though model complexity remained a bottleneck. Lin et al. expanded on this idea by utilizing Directed Statistics Pooling[5] , in order to increase the model's sensitivity to minute frequency patterns. However, these models are less appropriate for real-time or low-resource contexts since they frequently require extensive pretraining and big memory footprints.

Lu et al. proposed a One-Class Knowledge Distillation (OCKD)[6] model, which used a teacher-student architecture to generalize spoof detection using only genuine speech samples. Although layer-conditioned embedding fusion[7] was presented by Sinha et al. to improve generalization, real-time processing efficiency was still lacking.

Usage of Mel-Frequency Cepstral Coefficients (MFCC), Linear Frequency Cepstral Coefficients (LFCC), and Bark Frequency Cepstral Coefficients (BFCC)[8] has been prevalent. Power-Normalized Cepstral Coefficients (PNCC) and Constant-Q Cepstral Coefficients (CQCC) have also demonstrated potential in feature extraction. Zhang et al. introduced a dual-band fusion technique[9] that independently assesses low- and high-frequency spectrograms. This approach had overfitting and errors in Voice Activity Detection (VAD) during silent segments, despite the fact that it successfully addressed certain attack types.

Most existing models are either computationally demanding or largely reliant on spoofed samples. Our work addresses these challenges by presenting a lightweight, ResNet-based one-class learning model that solely uses real samples for training. Our system balances robustness, computational efficiency, and deployment readiness by integrating DocSoftmax-enhanced classification with spectral diversity (MFCC, LFCC, BFCC, and PQCC).

METHODS

The methods section discusses the tools, datasets, hardware, and software environments employed to develop and test the voice spoofing detection system.

DATA PREPROCESSING

This work uses the ASVspoof 2019 dataset as its main dataset, which includes both real and fake speech samples. The dataset includes a variety of spoofing attacks that cover different attack scenarios, such as voice conversion, replay attacks, and text-to-speech synthesis. To guarantee consistency in the frequency domain across all samples, all audio files were first resampled to a standard sampling rate of 16 kHz. In order to prevent problems with variable-duration inputs, format conversion was also carried out to standardize audio encoding, and trimming or padding was used to enforce a fixed length. In order to reduce distortions and reverberation effects, adaptive filtering was employed, while spectral gating was used to remove stationary background noise while preserving speech integrity. By broadening the variety of training samples, data augmentation was essential in enhancing generalization. Pitch shifting, time-stretching, and additive noise injection were among the augmentation techniques used. Additionally, controlled noise injection improved the model's adaptability to various recording environments by simulating real-world conditions. Z-score normalization and Min-Max scaling were used to standardize feature distributions, preventing the model from being biased during training by variations in intensity levels. Additionally, to keep the loudness of the various audio samples balanced, energy-based normalization was applied.

METHODOLOGY

Using a systematic methodology, the suggested voice spoofing detection system distinguishes between real and fake voice samples by utilizing one-class learning. The approach is divided into four major phases: To increase model generalization, the process begins with data collection and preprocessing, which involves standardizing the audio data, enhancing its quality, and applying data augmentation techniques. Then, feature extraction is performed by deriving important spectral features such as MFCC, LFCC, and BFCC to effectively differentiate between spoofed and genuine voices. In the next stage, a ResNet-based deep learning model is implemented and trained using a one-class learning approach to detect anomalies in speech patterns. Finally, the model's performance is evaluated using metrics such as Equal Error Rate (EER), F1-score, and overall accuracy, supported by comparative analysis and visualizations to validate its effectiveness.

A.FEATURE EXTRACTION

One of the most important processes in transforming unstructured audio signals into a representation fit for deep learning was feature extraction. The spectral and temporal aspects of speech were recorded using a variety of cepstral techniques. Time-domain signals were transformed into time-frequency representations using the Short-Time Fourier Transform (STFT). Mel Frequency Cepstral Coefficients (MFCCs) were extracted using the Mel scale and were computed as

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

where S_m represents Mel-filterbank energies. Additionally, LFCC, BFCC, CQCC, PNCC, and RPLP were extracted to enhance feature diversity. Delta and Delta-Delta coefficients were computed to capture speech dynamics. As part of the preprocessing pipeline, audio was resampled to 16 kHz, noise was reduced using spectral gating, and time-frequency representations were created using the Short-Time Fourier Transform. To ensure consistency across samples, extracted features were normalized using Z-score normalization to remove amplitude variations.

B.MODEL TRAINING

The proposed approach uses a deep learning architecture based on ResNet to detect voice spoofing. The reason ResNet, or Residual Neural Network, was selected is that it can efficiently learn deep feature representations without running into the vanishing gradient issue that plagues conventional deep CNNs. ResNet uses skip connections, which enable the gradient to propagate through the network efficiently, enabling deeper architectures without performance degradation, in contrast to traditional CNNs that might find it difficult to capture long-term dependencies in speech data. This is especially helpful for voice spoofing detection, where complex feature extraction is needed to extract fine-grained differences between real and spoof speech. Even though other architectures, like LSTMs and GRUs, are frequently employed for sequential speech processing, they can have trouble handling big datasets and frequently demand a lot of processing power. ResNet is the best option for this task because of its architecture, which allows for both local and deep hierarchical feature learning. Each of the several essential parts of the model architecture is intended to accurately classify and extract significant features from speech. Preprocessed speech features are sent to the input layer in the form of spectrogram-like representations. The working of the ResNet model is shown in the figure *3.1 RESNET Architecture* as the features are extracted and then fed into the ResNet model, which uses convolutional layers with specific filters to learn complex speech frequency representations. In order to mitigate vanishing gradient issues and facilitate deeper learning, residual blocks in ResNet enable gradients to flow effectively. Using a one-class learning approach, the model treats deviations as possible spoof attacks and focuses on learning the features of real speech.

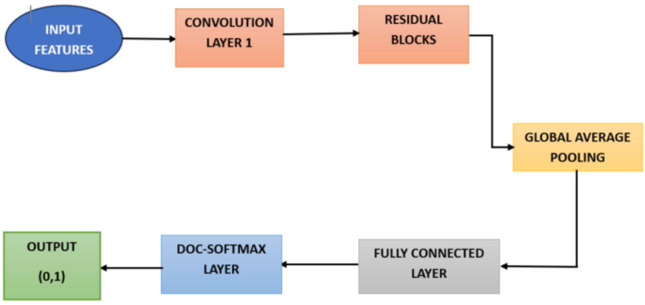


Fig.3.1 RESNET ARCHITECTURE

Multiple loss functions are included in order to address class imbalance and enhance generalization effectiveness. Class-Balanced Loss dynamically modifies each class's contribution to avoid overshadowing spoofed samples, which are generally underrepresented. Furthermore, class-dependent decision boundaries are applied by Label-Distribution-Aware Margin Loss to guarantee that spoof samples are not dominated by the real class. Using the Adam optimizer, the model is trained at a learning rate of 0.0001, utilizing early stopping to stop training when optimal performance is achieved, dropout layers to prevent overfitting, and batch normalization to stabilize learning. A DocSoftmax activation function that fine-tunes probability distributions and guarantees confident predictions comes after a fully connected layer to accomplish the final classification. Since the output is binary, "0" denotes real speech and "1" denotes spoof speech.

C.ANOMALY DETECTION

Anomaly detection is a crucial phase in the voice spoofing detection system. The training model was used to classify incoming audio samples throughout the detection phase. The sample was marked as genuine if the estimated likelihood exceeded an established threshold; if not, it was labeled as a spoof. Real-time predictions generated by user input have been rendered attainable by integrating the output into a Flask API. Once the audio file has been transmitted, the system examines it, gathers features, and predicts whether the voice is real or spoofed.

SYSTEM ARCHITECTURE

In order to accurately classify real and spoof speech, the suggested voice spoofing detection system's architecture uses a structured pipeline that integrates several modules. The Architecture diagram as seen in *Fig.4.1 Architecture Of Voice Spoofing Detection*, shows how data moves through the system where preprocessing is the first step, in which raw audio is segmented, noise-reduced, and normalized to ensure consistency across all inputs. Following that, feature extraction is carried out using spectrogram transformations, features to capture both perceptual audio features that resemble human hearing and fine-grained spectral details.

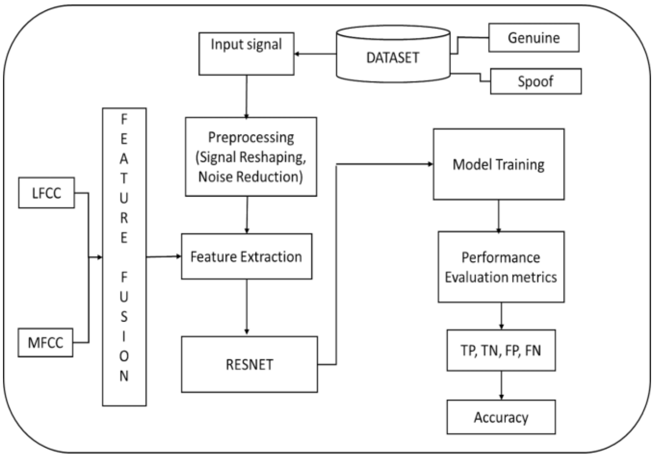
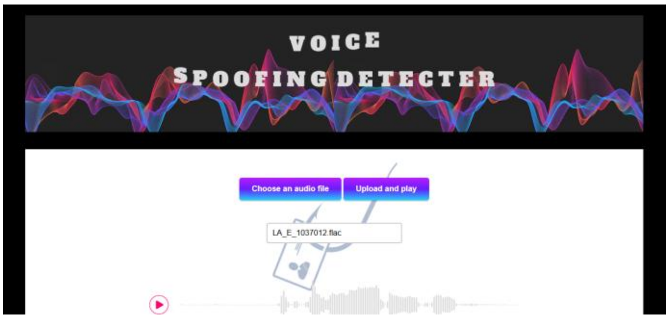


FIG.4.1 ARCHITECTURE OF VOICE SPOOFING DETECTION SYSTEM

For classification, the system employs a CNN architecture based on ResNet. The one-class learning model identifies outliers that may represent spoofing attempts and focuses on identifying complex patterns in the real world. The overall system workflow as seen in the user interface begins by providing the raw audio data as input, which is then preprocessed to standardize its format for consistent analysis. Then the essential audio features such as Linear Frequency Cepstral Coefficients (LFCC), Mel Frequency Cepstral Coefficients (MFCC), and spectrogram transformations are extracted. A ResNet-based one-class learning model that is trained solely on real voice samples uses these features as input. The model successfully separates real voices from possible spoof attempts after being trained and tested on previously unseen audio samples to evaluate and forecast their authenticity.

RESULTS

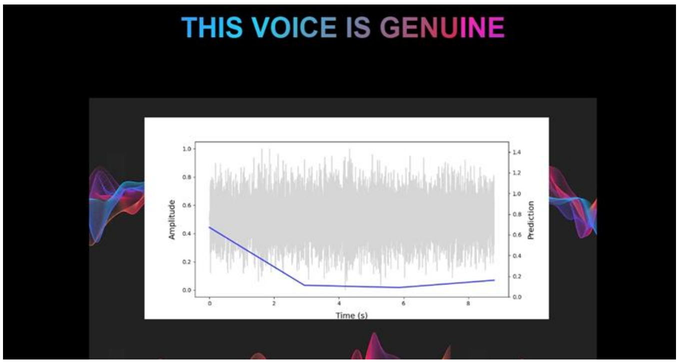
The purpose of this voice spoofing detection is to identify and prevent unwanted access by spotting fraudulent voices. The system gathers high-quality genuine voice samples, refines them, and extracts unique characteristics for model training. The model solely employs a one-class learning algorithm to learn genuine voice sample patterns to detect anomalies that may indicate spoofing attempts. Using a simple-to-use interface as seen in Fig.5.1, users may upload samples for instantaneous evaluation, making voice authentication readily available and invincible to spoofing techniques.



LA_E_1037012.flac

Fig.5.1 UI 1

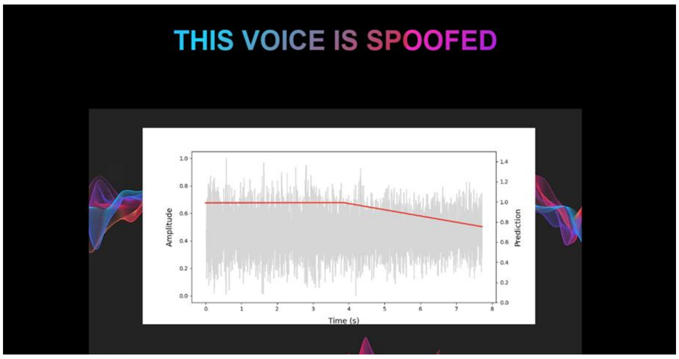
Anomalies are marked as spoofed, and real voices are marked as genuine as seen in Fig.5.2.



THIS VOICE IS GENUINE

Fig.5.2 OUTPUT 1

If the uploaded audio input is a spoofed voice, the output will be as shown in the figure 5.3. The binary range calculation for both samples is 0 for genuine and 1 for spoofed and the scalar range is -1 for spoofed voices and 1 for genuine voice samples.

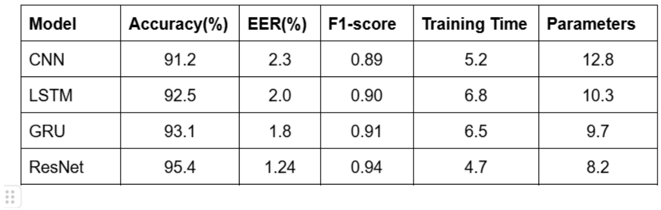


THIS VOICE IS SPOOFED

Fig.5.3 Output 2

The accuracy, equal error rate (EER), F1-score, and training time efficiency of the suggested ResNet-based voice spoofing detection system were the main metrics used for evaluation. The outcomes demonstrate that ResNet performs better at identifying spoofed voices while preserving computational efficiency than more conventional models like CNN, LSTM, and GRU.

Table.5.1 Performance Comparison



Model	Accuracy(%)	EER(%)	F1-score	Training Time	Parameters
CNN	91.2	2.3	0.89	5.2	12.8
LSTM	92.5	2.0	0.90	6.8	10.3
GRU	93.1	1.8	0.91	6.5	9.7
ResNet	95.4	1.24	0.94	4.7	8.2

A comparison of various models is shown in Table 5.1. The suggested system outperforms other architectures with an accuracy of 95.4%, an EER of 1.24%, and an F1-score of 0.94. ResNet has the quickest training time (4.7 hours) and uses fewer parameters despite its depth. ResNet effectively captures complex patterns in voice signals by preventing vanishing gradients, which is not possible with CNN or LSTMS.

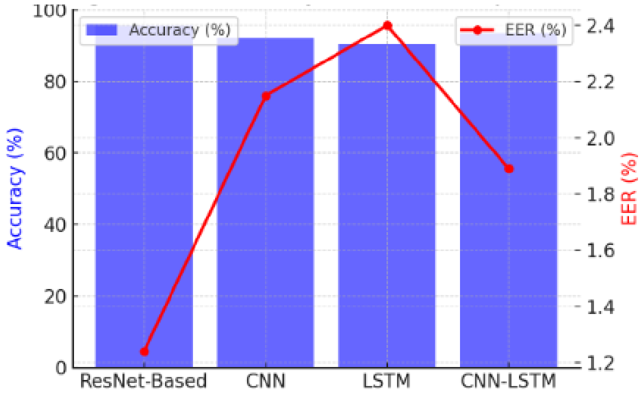


Fig.5.4 Accuracy and EER Comparison

A bar graph comparing Accuracy (%) and Equal Error Rate (EER %) across different models, as shown in Fig.5.4 demonstrates that the ResNet-based one-class model outperforms CNN, LSTM, and other architectures.

A line graph visualizing the F1-score comparison as shown in Fig.5.5, illustrates how the proposed model achieves higher precision and recall balance compared to previous techniques.

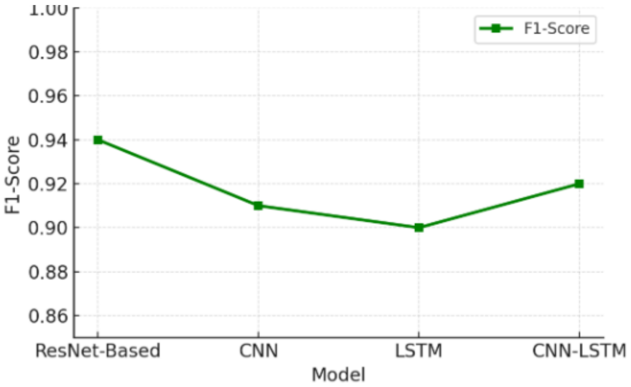


Fig.5.5 F1-score comparison

DISCUSSIONS

The proposed voice spoofing detection system is built on a one-class learning framework, a specialized machine learning technique designed to differentiate between real and spoofed audio recordings. This methodology is particularly suited for security-focused applications, as it enables the system to effectively learn from a dataset comprising only authentic (genuine) voice samples, without requiring labeled spoof data for training. By training exclusively on real voice data, the system becomes highly adept at identifying anomalies or deviations from natural speech patterns, making it particularly sensitive to spoofed audio.

The system showed strong detection capability and robustness by attaining low EER(1.24%) and high accuracy(95.4%). Security for voice authentication was significantly enhanced by combining spectral features with a ResNet-based framework, which improved detection performance in contrast to previous studies that discovered weaker generalizations and higher EERs (>2%). ResNet's highly computational nature restricts its use on devices with only a few resources. Adversarial training, edge device optimization, and developing datasets for different circumstances should be the main areas of future research.

CONCLUSION

This system successfully addresses the growing threat of voice spoofing in authentication systems by developing a lightweight and efficient detection model. By leveraging a ResNet-based student model with knowledge distillation and a one-class learning approach, the system can detect spoofed voices with high accuracy using only genuine voice samples for training. The integration of feature extraction, along with advanced noise reduction techniques, ensures robustness in diverse environmental conditions. The results demonstrate that our approach provides a scalable, adaptable, and resource-efficient solution for securing voice authentication in critical sectors like banking, healthcare, and government.Performance validation was conducted using the ASVspoof dataset, evaluating key metrics such as accuracy, recall, and F1-score to demonstrate the model's effectiveness. Future enhancements of this work is to implement self-learning capabilities that allow the system to update and refine its detection model over time, adapting to emerging spoofing techniques without frequent manual retraining and also to optimize the model for deployment on mobile and IoT-based authentication systems, ensuring efficient real-time processing on resource-constrained devices.

REFERENCES

[1] M. Todisco, H. Delgado, and N. W. D. Evans, "Constant Q Cepstral Coefficients: A Spoofing Countermeasure for Automatic Speaker Verification," *Computer Speech & Language*, vol. 45, pp. 516–535, 2017.

[2] T. Kinnunen et al., "ASVspoof 2019: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan," *Proc. Interspeech*, 2019.

[3] A. Lavrentyeva et al., "STC Anti-spoofing Systems for the ASVspoof 2019 Challenge," *Proc. Interspeech*, 2019.

[4] X. Xia, J. Yamagishi, and T. Kinnunen, "Squeeze-Excitation Networks for Voice Spoofing Detection," *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2020.

5] G. Lin, D. Luo, and J. Huang, "One-Class Neural Network with Directed Statistics Pooling for Spoofing Speech Detection," *IEEE Transactions on Information Forensics and Security*, vol. XX, pp. XX–XX, 2024.

[6] J. Lu, Y. Zhang, W. Wang, Z. Shang, and P. Zhang, "One-Class Knowledge Distillation for Spoofing Speech Detection," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11251–11255.

[7] S. Sinha, S. Dey, and G. Saha, "Improving Self-Supervised Learning Model for Audio Spoofing Detection with Layer-Conditioned Embedding Fusion," *Computer Speech & Language*, vol. 86, Article 101599, 2024.

[8] J. Li, H. Wang, P. He, S. M. Abdullahi, and B. Li, "Long-Term Variable Q Transform: A Novel Time-Frequency Transform Algorithm for Synthetic Speech Detection," *Digital Signal Processing*, vol. 120, Article 103256, 2022.

[9] Y. Zhang, W. Wang, and P. Zhang, "The Effect of Silence and Dual-Band Fusion in the Anti-Spoofing System," *Proc. Interspeech*, 2021, pp. 4279–4283.

[10] H. Delgado et al., "ASVspoof 2017 Version 2.0: Evaluation Campaign for the Detection of Replay Attacks," *Proc. Interspeech*, 2017.