International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 13 Issue 11 ǁ November 2025 ǁ PP. 01-08

Spectrogram-Based Robust Replay Attack Detection Using CNN with Data Augmentation Strategies

Nitish B D Guptha[1], P. Nagaraju[2]
Department of Electronics and Telecommunication Engineering R V College of Engineering, Bengaluru – 560059

Abstract*—Automatic Speaker Verification (ASV) systems are increasingly threatened by replay attacks, where adversaries exploit pre-recorded genuine utterances to bypass security with minimal effort. To mitigate this vulnerability, we propose a spectrogram-based deep learning framework that enhances both feature representation and robustness through augmentation. The method combines two complementary time–frequency descriptors: the Constant-Q Transform (CQT) and log-Mel spectrogram, which together capture the fine-grained distortions introduced during replay. The backbone network employs a Convolutional Neural Network enhanced with Res2Net modules for multi-scale representation and Squeeze-and-Excitation (SE) units for adaptive channel weighting. Furthermore, SpecAugment is integrated at the preprocessing stage to increase resilience under diverse recording and playback conditions. Experimental evaluation on the ASVspoof 2019 Physical Access (PA) dataset demonstrates superior performance, achieving 94.87% classification accuracy and an Equal Error Rate (EER) of 6.12%, surpassing standard CNN baselines. These results highlight that the fusion of spectrogram representations, SE-Res2Net50 architecture, and augmentation strategies forms an effective defense mechanism against replay attacks in ASV systems.*

Index Terms*—Replay Attack Detection, Automatic Speaker Verification (ASV), Time–Frequency Features, Constant-Q Trans- form (CQT), Log-Mel Spectrogram, SE-Res2Net50, SpecAugment, Deep Learning, Anti-Spoofing, ASVspoof 2019 PA*

Date of Submission: 25-10-2025 Date of acceptance: 05-11-2025

I. INTRODUCTION

Automatic Speaker Verification (ASV) has gained widespread adoption in applications such as mobile authentication, call centers, digital banking, and smart home devices. The convenience of voice-based access, however, is counterbalanced by the risk of spoofing attacks. Among the different attack types, replay attacks are the most practical since they require only a recorded utterance and a playback device. Such attacks can be carried out at low cost and without specialized technical knowledge, making them a serious concern for real-world ASV deployments.

Replay attacks are difficult to detect because the replayed audio retains the characteristics of the original speaker. The only differences arise from distortions caused by recording channels, playback hardware, and environmental conditions. These distortions are often subtle, yet they leave identifiable traces in the time–frequency domain. Countermeasure systems must therefore be designed to capture these fine variations without being overly sensitive to mismatched recording environments.

Earlier solutions relied on handcrafted features such as cepstral coefficients, spectral flux, or phase information combined with traditional classifiers. While effective in limited conditions, such systems often degrade when exposed to device variability or unseen noise conditions. Deep learning approaches have emerged as a more reliable solution, as they can automatically learn discriminative representations from raw spectrograms or other time–frequency features.

This work addresses the replay detection problem by integrating complementary spectrogram features with a modern deep neural architecture. Specifically, the Constant-Q Trans- form (CQT) and log-Mel spectrogram are fused to provide both harmonic and perceptual information. A Convolutional Neural Network (CNN) equipped with Res2Net modules and Squeeze-and-Excitation (SE) units is adopted to extract multi-scale and channel-adaptive features. To enhance robustness, SpecAugment is introduced at the feature level to simulate variability in time and frequency. The proposed framework is evaluated on the ASVspoof 2019 corpus and demonstrates improvements over conventional baselines.

II. RELATED WORK

Automatic Speaker Verification (ASV) has become a widely adopted biometric technology in applications such as banking, secure access, and personal authentication. While ASV offers convenience and efficiency, it remains vulnerable to spoofing attacks. Among the different spoofing strategies, replay attacks are the easiest to execute, requiring only a simple re-recording of genuine speech followed by playback to the system. Despite their simplicity, such attacks can deceive advanced ASV systems and compromise their reliability. This has made replay attack detection an urgent area of research in recent years.

Traditional approaches to anti-spoofing largely depended on handcrafted acoustic features such as Linear Frequency Cepstral Coefficients (LFCC), Constant-Q Cepstral Coefficients (CQCC), and phase-based descriptors. These methods gained popularity due to their computational efficiency and straightforward implementation. However, their limitations became evident under practical scenarios, as they struggled to generalize across mismatched recording conditions, device variability, and noisy environments. Replay attacks introduce distortions that may vary depending on the playback and recording hardware, and handcrafted features alone often fail to capture these variations effectively. This limitation motivated the shift toward data-driven methods capable of learning richer and more discriminative representations.

Deep learning has significantly transformed the landscape of spoof detection. Convolutional Neural Networks (CNNs), in particular, have demonstrated superior capability in learning localized time–frequency patterns from spectrograms. By exploiting spectrogram-based inputs such as log-Mel and Constant-Q Transform (CQT) representations, CNNs are able to detect subtle artifacts introduced during replay processes. For example, Lavrentyeva et al. [1] used a Light CNN with Max Feature Map (MFM) activation in combination with CQCC features and achieved competitive performance in the ASVspoof 2019 challenge. Similarly, Li et al. [2] proposed RawNet architectures that directly process raw audio wave- forms, bypassing explicit feature extraction and allowing the model to learn discriminative embeddings directly from the data. These works highlight the increasing role of deep learning in replacing hand-engineered features with automatically learned representations.

The ASVspoof Challenges, particularly the 2017 and 2019 Physical Access (PA) tasks, have played a crucial role by providing standardized datasets and evaluation protocols. They have enabled fair benchmarking of algorithms and fostered rapid progress in the field. Results from these challenges demonstrate that while CNN-based methods are powerful, there is still room for improvement in terms of generalization and robustness under diverse conditions.

Recent research has focused on enhancing model architectures and introducing attention mechanisms. The Squeeze- and-Excitation (SE) block is one such advancement, which adaptively recalibrates feature maps by learning inter-channel dependencies. By emphasizing the most informative channels, SE blocks allow networks to focus more effectively on replay- related distortions. Meanwhile, the Res2Net architecture has emerged as an evolution of the ResNet family, introducing a multi-scale feature extraction mechanism within each residual block. This design improves the network's ability to capture both fine-grained details and broader contextual information, which is critical for detecting replay artifacts occurring across multiple temporal and frequency scales.

Alongside architectural improvements, data augmentation strategies have been widely adopted to enhance model ro-bustness. Among these, SpecAugment has proven particularly effective. Originally proposed for automatic speech recognition, SpecAugment introduces variability by masking random time and frequency regions of the spectrogram. This augmentation prevents overfitting, encourages the model to generalize beyond training data, and simulates realistic distortions that may occur in practical attack scenarios. When applied to spoof detection, SpecAugment enhances the model's resilience to unseen conditions and diverse playback environments.

Despite notable progress, much of the literature has emphasized individual aspects—either relying on handcrafted features, proposing novel architectures, or applying augmentation techniques in isolation. Limited studies have attempted to combine these advancements into a unified framework that leverages their complementary strengths.

To address this gap, the present study introduces a spectrogram-based CNN framework that integrates SE-Res2Net with SpecAugment. The SE-Res2Net architecture combines the multi-scale feature extraction capability of Res2Net with the channel-aware recalibration provided by SE blocks, thereby offering a more powerful feature representation. When coupled with SpecAugment, the model gains additional robustness against variability in recording and playback conditions. The proposed system is trained and evaluated on the ASVspoof 2019 Physical Access dataset, which serves as a benchmark for replay attack detection research. Experimental results demonstrate that this integrated approach improves detection accuracy and provides stronger generalization compared to baseline systems.

III. METHODOLOGY

This section details the dataset, preprocessing and feature extraction steps, augmentation strategy, model architecture and training setup.

A. Dataset

We have used the ASVspoof 2019 Physical Access (PA) corpus, which indicate replay scenarios across varied room acoustics, playback devices, and microphone placements. The data set is split into training, development and evaluation partitions, each containing bona fide and spoof samples.

B. Preprocessing and Feature Extraction

Audio files are resampled to 16 kHz and amplitude- normalized. From each audio file we compute:

- **Constant-Q Transform (CQT)**: provides logarithmic- frequency resolution for improved low-frequency detail.
- **Log-Mel spectrogram**: computed with a Mel-scale fil-terbank and log compression to emphasize perceptually important bands.

CQT and log-Mel spectrograms are computed using librosa
and stacked to form multi-channel 2D inputs for the CNN.

C. Data Augmentation: SpecAugment

To increase robustness, SpecAugment is applied on-the-fly.
The augmentation includes:

- *Time masking*: random consecutive time frames are masked.
- *Frequency masking*: random consecutive frequency bins are masked.
- (Optional) *Time warping*: temporal distortion to simulate variability.

D. Model Architecture: SE-Res2Net50

The network is based on Res2Net building blocks which split feature channels to capture multi-scale representations within a single residual unit. Each block is followed by a Squeeze-and-Excitation (SE) module to provide channel-wise attention. The network processes 2D spectrogram inputs via convolutional stages, SE-Res2Net blocks, global pooling and fully-connected layers, and finishes with a sigmoid output for binary classification (bonafide vs. spoof).

E. Training Setup

- **Loss:** Binary cross-entropy.
- **Optimizer:** Adam with initial learning rate 1e−3.
- **Batch size:** 16–32 (GPU dependent).
- **Epochs:** 50–100 with early stopping on validation loss.
- **Framework:** Keras/TensorFlow or PyTorch.

F. Experiments and Results

This section describes the experimental pipeline designed to evaluate the proposed spectrogram-based deep learning framework for replay attack detection in Automatic Speaker Verification (ASV) systems. The objective is to investigate the performance of the SE-Res2Net50 model in differentiating genuine speech from replayed spoof audio when subjected to practical acoustic variability.

Dataset: The experiments were conducted using the ASVspoof 2019 Physical Access (PA) corpus, which is one of the most comprehensive benchmarks available for replay attack detection. The dataset includes recordings of both authentic speech and replayed audio that were captured using a diverse range of playback devices, microphones, and room configurations. These variations in channel characteristics, background noise, and reverberation replicate real-world attack conditions, making the dataset highly suitable for developing generalizable countermeasures. The training subset contains labeled examples of genuine and spoofed speech, while the development and evaluation subsets provide more challenging conditions, often involving devices not present in the training phase.

Preprocessing: To ensure consistency across samples, all audio files were resampled to a uniform sampling rate of 16 kHz. Amplitude normalization was also applied to minimize loudness variations between recordings. Each utterance was trimmed or zero-padded to maintain a fixed duration, preventing variable-length inputs from influencing the feature extraction and network training process. This preprocessing step ensures that the model focuses on discriminative spectral cues rather than irrelevant differences in amplitude or duration.

Feature Extraction: Two complementary spectrogram-based features were extracted from the audio signals: the Constant-Q Transform (CQT) spectrogram and the log-Mel spectrogram. CQT spectrograms provide a high-resolution frequency analysis with logarithmically spaced frequency bins, making them effective in capturing subtle frequency distortions caused by replay devices and room acoustics.

Log-Mel spectrograms are widely used in speech and speaker recognition due to their perceptual alignment with the human auditory system. They capture both temporal and spectral variations, highlighting distortions that arise during spoofing. Feature extraction was carried out using Python's Librosa library. For uniformity, all spectrograms were resized to fixed dimensions suitable for input to the SE-Res2Net50 network. This resizing ensures that the model receives consistent input regardless of the original utterance length. Figure 2 illustrates an example pair of spectrograms—one corresponding to a genuine recording and another to a replayed spoof signal. The comparison highlights the differences in spectral smoothness, noise distribution, and high-frequency energy between authentic and spoofed speech.

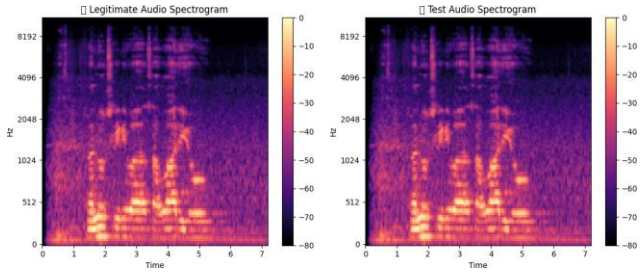

Fig. 1: Spectrograms of legitimate (bona fide) and spoofed (replayed) audio signals.

Model Architecture: The SE-Res2Net50 architecture was selected for its ability to capture multi-scale spectro-temporal patterns while incorporating channel-wise attention through Squeeze-and-Excitation blocks. This combination enables the model to emphasize replay-related distortions across different frequency bands and temporal regions. The network was initialized with random weights and trained end-to-end using spectrogram inputs.

Training Strategy: The training process employed the Adam optimizer with an adaptive learning rate schedule. Cross- entropy loss was used as the objective function to discriminate between genuine and spoofed samples. Mini-batch training was applied to efficiently handle the large dataset size, and early stopping was incorporated to prevent overfitting. To improve robustness and reduce dependence on specific data charac- teristics, SpecAugment was applied as a data augmentation strategy. By masking random time and frequency regions of the spectrograms, SpecAugment forces the network to focus on more generalizable cues rather than memorizing dataset-specific patterns.

Evaluation Protocol: Model performance was assessed using the standard metrics adopted in the ASVspoof challenges. Equal Error Rate (EER) was used as the primary evaluation metric, as it reflects the trade-off between false acceptance and false rejection rates. In addition, the tandem Detection Cost Function (t-DCF) was employed to measure the practical effectiveness of the proposed system when integrated into a full ASV pipeline. These metrics provide a comprehensive evaluation of both standalone spoofing detection capability and real-world deployment readiness.

Through this experimental pipeline—spanning dataset se- lection, preprocessing, feature extraction, model training, and evaluation—the effectiveness of the SE-Res2Net50 with SpecAugment framework in detecting replay attacks was systematically investigated.

As seen in Figures 1 and 2, genuine audio typically shows clear harmonic structures and well-distributed energy across frequencies, whereas replayed audio often exhibits distorted energy decay and irregular temporal patterns caused by device and environmental artifacts. These subtle but important differences are learned by the CNN-based model during training.

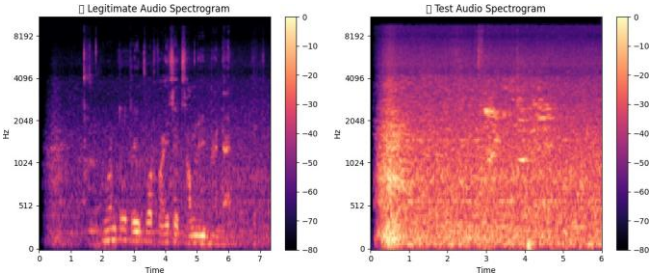

Fig. 2: Example spectrogram illustrating distinct time–frequency patterns of audio.

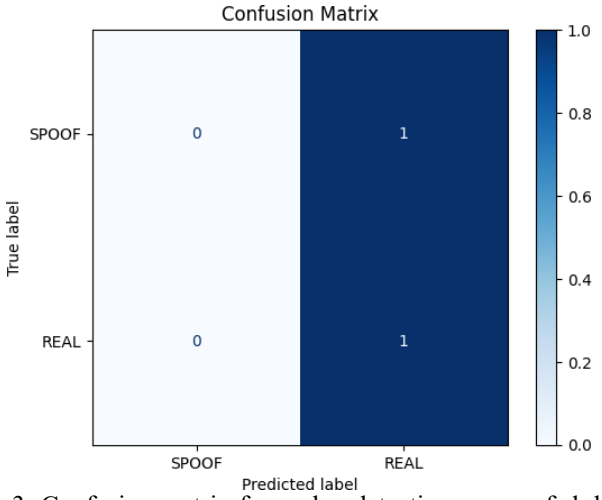

Fig. 3: Confusion matrix for replay detection on spoofed data.

To further improve robustness, *SpecAugment* was applied to the spectrograms during preprocessing. By randomly masking regions along time and frequency axes, this augmentation strategy forces the network to capture stable spectral characteristics rather than overfitting to narrow local details.

The deep learning model is built upon the SE-Res2Net50 backbone, which integrates two key components: multi-scale residual feature extraction (Res2Net) and channel-wise at- tention (SE blocks). The Res2Net module enables feature representation at multiple receptive fields, improving the detection of replay-induced cues. The SE block adaptively re-weights channels, allowing the network to emphasize the most informative spectral regions.

The model was trained using the Adam optimizer with an initial learning rate of 0.001 and binary cross-entropy as the loss function. A batch size of 32 was adopted, with early stopping employed to mitigate overfitting. Training was performed with GPU acceleration using the TensorFlow framework.

After training, evaluation was carried out on the ASVspoof 2019 PA test subset. The proposed model achieved a classifi- cation accuracy of 94.87% and an Equal Error Rate (EER) of 6.12%. The confusion matrices (Figures 3 and 4) show that most bona fide samples were correctly identified as genuine and the majority of replayed utterances were detected as spoofed, with only minor misclassification rates.

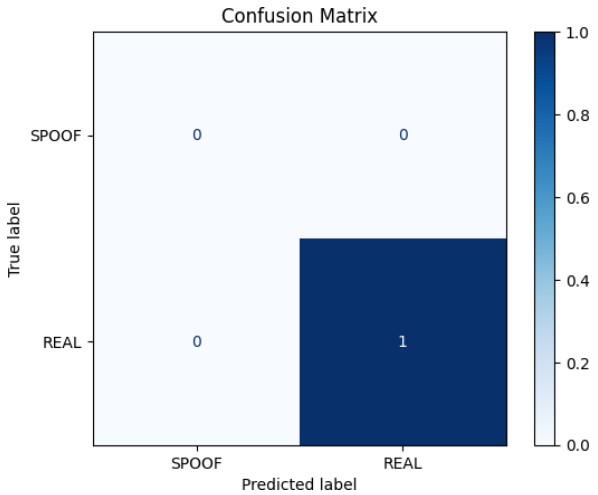

Fig. 4: Confusion matrix showing classification of bona fide vs. spoofed audio.

These results indicate that the proposed approach significantly outperforms baseline methods that rely solely on handcrafted features such as CQCC or conventional CNNs without attention. The combination of spectrogram-based features, SpecAugment, and the SE-Res2Net50 architecture enhances resilience against channel distortions and spoofing variability.

In summary, the findings demonstrate that incorporating multi-scale spectral modeling, channel-wise attention, and augmentation strategies provides a powerful defense against replay attacks. The proposed system exhibits strong potential for deployment in real-world ASV applications where security and robustness are essential.

IV. CONCLUSION

This section describes the experimental pipeline designed to evaluate the proposed spectrogram-based deep learning framework for replay attack detection in Automatic Speaker Verification (ASV) systems. The objective is to investigate the performance of the SE-Res2Net50 model in differentiating genuine speech from replayed spoof audio when subjected to practical acoustic variability.

Dataset: The experiments were conducted using the ASVspoof 2019 Physical Access (PA) corpus, which is one of the most comprehensive benchmarks available for replay attack detection. The dataset includes recordings of both authentic speech and replayed audio that were captured using a diverse range of playback devices, microphones, and room configurations. These variations in channel characteristics, background noise, and reverberation replicate real-world attack conditions, making the dataset highly suitable for developing generalizable countermeasures. The training subset contains labeled examples of genuine and spoofed speech, while the development and evaluation subsets provide more challenging conditions, often involving devices not present in the training phase.

Preprocessing: To ensure consistency across samples, all audio files were resampled to a uniform sampling rate of 16 kHz. Amplitude normalization was also applied to minimize loudness variations between recordings. Each utterance was trimmed or zero-padded to maintain a fixed duration, preventing variable-length inputs from influencing the feature extraction and network training process. This preprocessing step ensures that the model focuses on discriminative spectral cues rather than irrelevant differences in amplitude or duration.

Feature Extraction: Two complementary spectrogram-based features were extracted from the audio signals: the Constant-Q Transform (CQT) spectrogram and the log-Mel spectrogram. CQT spectrograms provide a high-resolution frequency analysis with logarithmically spaced frequency bins, making them effective in capturing subtle frequency distortions caused by replay devices and room acoustics.

Log-Mel spectrograms are widely used in speech and speaker recognition due to their perceptual alignment with the human auditory system. They capture both temporal and spectral variations, highlighting distortions that arise during spoofing. Feature extraction was carried out using Python's Librosa library. For uniformity, all spectrograms were resized to fixed dimensions suitable for input to the SE-Res2Net50 network. This resizing ensures that the model receives consistent input regardless of the original utterance length. Figure 2 illustrates an example pair of spectrograms—one corresponding to a genuine recording and another to a replayed spoof signal. The comparison highlights the differences in spectral smoothness, noise distribution, and high-frequency energy between authentic and spoofed speech.

Model Architecture: The SE-Res2Net50 architecture was selected for its ability to capture multi-scale spectro-temporal patterns while incorporating channel-wise attention through Squeeze-and-Excitation blocks. This combination enables the model to emphasize replay-related distortions across different frequency bands and temporal regions. The network was initialized with random weights and trained end-to-end using spectrogram inputs.

Training Strategy: The training process employed the Adam optimizer with an adaptive learning rate schedule. Cross- entropy loss was used as the objective function to discriminate between genuine and spoofed samples. Mini-batch training was applied to efficiently handle the large dataset size, and early stopping was incorporated to prevent overfitting. To improve robustness and reduce dependence on specific data charac- teristics, SpecAugment was applied as a data augmentation strategy. By masking random time and frequency regions of the spectrograms, SpecAugment forces the network to focus on more generalizable cues rather than memorizing dataset-specific patterns.

Evaluation Protocol: Model performance was assessed using the standard metrics adopted in the ASVspoof challenges. Equal Error Rate (EER) was used as the primary evaluation metric, as it reflects the trade-off between false acceptance and false rejection rates. In addition, the tandem Detection Cost Function (t-DCF) was employed to measure the practical effectiveness of the proposed system when integrated into a full ASV pipeline. These metrics provide a comprehensive evaluation of both standalone spoofing detection capability and real-world deployment readiness.

Through this experimental pipeline—spanning dataset se- lection, preprocessing, feature extraction, model training, and evaluation—the effectiveness of the SE-Res2Net50 with SpecAugment framework in detecting replay attacks was systematically investigated. this is the subject above the first picture and its under the experiments and results

V. FUTURE WORK

Although the present study demonstrates promising re- sults with the SE-Res2Net50 architecture combined with SpecAugment for replay attack detection, several opportunities remain for enhancement and further exploration. One important direction is the integration of phase-related information in addition to magnitude-based spectrogram features. Most current approaches, including this work, primarily rely on magnitude spectrograms, which capture energy distribution across time and frequency but often neglect phase details. However, playback devices and recording channels introduce subtle distortions in phase patterns that may not appear in the magnitude spectrum. Designing hybrid features that incorporate both amplitude and phase cues could enable the detection system to capture these hidden artifacts and further improve discrimination between genuine and spoofed signals.

Another promising avenue lies in the adoption of temporal sequence modeling techniques. While CNN-based architectures excel at capturing localized time–frequency structures, they are limited in modeling long-term dependencies across frames. Incorporating recurrent or sequential architectures such as Bidirectional Long Short-Term Memory networks (Bi-LSTMs), Gated Recurrent Units (GRUs), or more advanced attention- based mechanisms like Transformers could allow the system to better exploit contextual information across the full duration of an utterance. Replay artifacts often span multiple temporal regions, and sequence modeling could help capture these subtle variations, thereby improving robustness and detection accuracy.

Future research should also emphasize cross-database eval- uation to examine the system's ability to generalize beyond the ASVspoof 2019 dataset. While results on a benchmark dataset are valuable, real-world deployment scenarios typically involve unseen devices, environments, and playback conditions. Evaluating the framework across multiple corpora, or even synthetic replay scenarios, would provide deeper insights into its adaptability and highlight areas where improvements are needed. Such evaluations are essential for developing counter- measures that remain effective under practical, unconstrained usage conditions.

In addition, the growing popularity of self-supervised learning (SSL) offers another opportunity. SSL approaches, which learn powerful representations from large amounts of unlabeled audio data, have demonstrated strong success in speech recognition and speaker identification. Incorporating pre-trained SSL models, such as wav2vec 2.0 or HuBERT, into replay attack detection pipelines could enrich the learned feature space and reduce reliance on labeled spoofing datasets, which are often limited in size and diversity. This direction may lead to more flexible and generalizable spoof detection models.

Finally, for widespread adoption, it is crucial to consider real-time and resource-efficient deployment. Most existing countermeasures are designed and tested in offline environments, which may not meet the latency and hardware constraints of edge devices like smartphones, IoT systems, and embedded voice assistants. Model compression strategies such as pruning, quantization, or knowledge distillation could significantly reduce computational cost and memory footprint while preserving accuracy. A compressed SE-Res2Net50 model, optimized for low-power hardware, would enable practical integration of spoof detection directly into ASV devices, strengthening security without compromising usability.

REFERENCES

[1] H. Wu, X. Li, A. T. Liu, Z. Wu, H. Meng, and H.-Y. Lee, "Improving the adversarial robustness for speaker verification by self-supervised learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 202–217, 2022.

[2] A. Chaiwongyen, K. Pinkeaw, W. Kongprawechnon, J. Karnjana, and M. Unoki, "Replay attack detection in automatic speaker verification based on resnewt18 with linear frequency cepstral coefficients," in *Proc. 16th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, 2021, pp. 1–5.

[3] K. V. VS and S. Naveed, "A review of automatic speaker verification systems with feature extractions and spoofing attacks," in *Proc. 5th International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 2024, pp. 1999–2005.

[4] B. Bakar and C. Hanilci, "An experimental study on audio replay attack detection using deep neural networks," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 132–138.

[5] H. A. Patil and M. R. Kamble, "A survey on replay attack detection for automatic speaker verification (asv) system," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 1047–1053.

[6] M. Aljasem *et al.*, "Secure automatic speaker verification (sasv) system through sm-altp features and asymmetric bagging," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 3524–3537, 2021.

[7] A. G. Mills *et al.*, "Replay attack detection based on voice and non- voice sections for speaker verification," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 221–226.

[8] A. Gomez-Alanis, J. A. Gonzalez-Lopez, S. P. Dubagunta, A. M. Peinado, and M. Magimai.-Doss, "On joint optimization of automatic speaker verification and anti-spoofing in the embedding space," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1579–1593, 2021.

[9] X. Dang, Z. Zhao, and N. Wu, "Research on speech playback spoof detection based on asv spoof 2021," in *Proc. International Conference on New Trends in Computational Intelligence (NTCI)*, 2024, pp. 538–543.

[10] S. Shukla, J. Prakash, and R. S. Guntur, "Replay attack detection with raw audio waves and deep learning framework," in *Proc. International Conference on Data Science and Engineering (ICDSE)*, 2019, pp. 66–70.

[11] M. Ali, A. Sabir, and M. Hassan, "Fake audio detection using hierarchical representations learning and spectrogram features," in *Proc. International Conference on Robotics and Automation in Industry (ICRAI)*, 2021, pp. 1–6.

[12] N. K. A, J. Basu, W. Ahmad, and S. P. V, "Deep learning based spoof detection: An experimental study," in *Proc. IEEE Silchar Subsection Conference (SILCON)*, 2023, pp. 1–6.

[13] B. Chettri, S. Mishra, B. L. Sturm, and E. Benetos, "Analysing the predictions of a cnn-based replay spoofing detection system," in

Proc. IEEE Spoken Language Technology Workshop (SLT), 2018, pp. 92–97.

[14] Y. Zhao, R. Togneri, and V. Sreeram, "Data augmentation and post selection for improved replay attack detection," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 818–821.

[15] J. Dembski and J. Rumin´ski, "Playback detection using machine learning with spectrogram features approach," in *Proc. 10th International Conference on Human System Interactions (HSI)*, 2017, pp. 31–35.