PeerJ Computer Science

DeepRawNet: empowering deepfake audio detection through dynamic enhancements

Lubna A. Alharbi[1], Jasim Alnahas[2], Ali Javed[3,4], A'aeshah Alhakamy[1], Marriam Nawaz[4], Hafiz Malik[5], Hafsa Ilyas[4] and Faris Alfaifi[2]

[1] Computer Science Department, University of Tabuk, Tabuk, Saudi Arabia
[2] Industrial Engineering Department, University of Tabuk, Tabuk, Saudi Arabia
[3] James Watt School of Engineering, University of Glasgow, Glasgow, United Kingdom
[4] Software Engineering Department, University of Engineering and Technology, Taxila, Punjab, Pakistan
[5] Electrical and Computer Engineering Department, University of Michigan, Dearborn, Michigan, United States

ABSTRACT

Background: The generation of deepfake audio poses significant challenges to the reliability and security of automatic speaker verification (ASV)-based systems. ASV systems having applications in fintech, surveillance, home automation, security, *etc.*, are susceptible to a variety of deepfake/voice cloning attacks, including speech synthesis and voice conversion (VC). Impostors launch audio deepfake attacks on ASV systems to compromise their security and cause financial losses, data breaches, *etc.*

Methods: To combat such threats, we propose a robust and generalized audio deepfakes detection framework, DeepRawNet, by processing raw audio waveforms. Specifically, DeepRawNet is the enhanced version of RawNet2, introduces three key innovations: (1) we employ the Parametric Rectified Linear Unit (PReLU) activation in the residual blocks over Leaky ReLU in the RawNet2, introducing a learnable negative slope to enhance adaptive feature extraction, (2) substituting the simple convolution layer with a transpose convolution layer in the residual block addresses downsampling issues while preserving fine-grained temporal information crucial for capturing complex patterns in raw audio, (3) we incorporate the LogSoftmax activation function to stabilize and optimize learning during training and inference. These architectural refinements empower our DeepRawNet model with improved adaptability, robust learning capabilities, and enhanced capacity to capture complex temporal dependencies and discriminative patterns in the audio, making it a more effective solution for audio deepfake detection.

Results: We performed a rigorous evaluation of our proposed method on ASVspoof2019-LA and ASVspoof2021-LA/DF datasets, including algorithm-wise and cross-corpora evaluation, an ablation study with different model configurations, and comparison against baseline models and existing approaches. Experimental results highlight the improved performance of DeepRawNet against the ASVspoof baselines, improved generalization across diverse spoofing attacks, particularly for the most challenging VC attacks, and effectiveness in combating deepfake audio threats.

Subjects Artificial Intelligence, Computer Vision, Data Mining and Machine Learning
Keywords Audio spoofing, Deepfake detection, Deep learning, Logical access attacks, Text-to-speech synthesis, Voice conversion

Submitted 9 June 2025
Accepted 15 January 2026
Published 12 March 2026

Corresponding author
Lubna A. Alharbi,
Lualharbi@ut.edu.sa

Academic editor
Ankit Vishnoi

Additional Information and Declarations can be found on page 28

DOI **10.7717/peerj-cs.3670**

How to cite this article Alharbi LA, Alnahas J, Javed A, Alhakamy A, Nawaz M, Malik H, Ilyas H, Alfaifi F. 2026. DeepRawNet: empowering deepfake audio detection through dynamic enhancements. **PeerJ Comput. Sci. 12:e3670 DOI 10.7717/peerj-cs.3670**

INTRODUCTION

Automatic Speaker Verification (ASV) stands as a pivotal technology within the realm of biometric authentication, binding the unique vocal characteristics of individuals to authenticate their identity (*Liz-López et al., 2024*; *Habbash et al., 2024*). ASV finds diverse applications across various sectors, contributing significantly to strengthening security measures. In access control, ASV systems enhance traditional methods by providing secure entry to restricted areas based on voice verification. Furthermore, the financial sector influences ASV for secure authorization of transactions. Call centers integrate ASV to authenticate users during customer service interactions, enhancing data security and preventing unauthorized access. Mobile devices incorporate ASV technology for user authentication, allowing individuals to unlock their smartphones or perform secure transactions through spoken passphrases (*Yang et al., 2024*). Additionally, ASV plays a crucial role in forensic investigations, aiding in the analysis of voice recordings for identification purposes. The technology contributes to voice biometrics, facilitating identity verification systems in scenarios where secure and non-intrusive authentication is paramount (*Hijji & Alam, 2021*). The field of ASV systems is increasingly challenged by the emergence of voice spoofing techniques, which aim to deceive the system's authentication processes. These techniques introduce significant vulnerabilities, affecting the reliability of ASV systems and compromising the security of voice-based identification (*Das, 2021*). These techniques are broadly categorized as: (i) Physical voice spoofing, which involves the use of pre-recorded audio samples to mimic the targeted individual's voice (*Khan et al., 2023b*), and (ii) Logical voice spoofing (audio deepfakes), *i.e.*, voice conversion and voice synthesis, where voice conversion specifically involves altering the characteristics of a speaker's voice to match a predefined target, while voice synthesis utilizes advanced voice synthesis generation algorithms to generate artificial voice samples that closely resemble the target's voice (*Khan et al., 2023a*). The fact that the converted voice originates from a real person, which contains vigorous variations of the human voice, unlike speech synthesis without such variations, makes voice conversion more challenging to detect.

Recently, we have observed the frequent use of smart speakers (SS) like Amazon's Echo and Google Home with built-in ASV technology in different application domains, including home automation, online banking, forensics, *etc.* We have physically tested the ASV capability of these SSs in our lab and found that they are unable to identify whether the given audio sample is bonafide or synthetic. This shows that the SSs are prone to a variety of audio spoofing attacks, including deepfakes. The easier availability of AI tools and modern-day generative algorithms has made it easy to create convincing synthetic audios that the impostors can use to spread disinformation campaigns (Fig. 1A), leading to political instability, chaos in financial markets, *etc.* Similarly, such audio deepfakes can be launched *via* SS to gain access to someone's online banking account, as shown in Fig. 1B, resulting in financial scams. Similarly, an imposter can create and play the deepfake voice to an SS to breach the security of someone's home, as demonstrated in Fig. 1C. One notable incident refers to financial fraud, where attackers exploited logical voice spoofing to fool ASV systems during transaction authorizations (*Dawood et al., 2022*). This incident

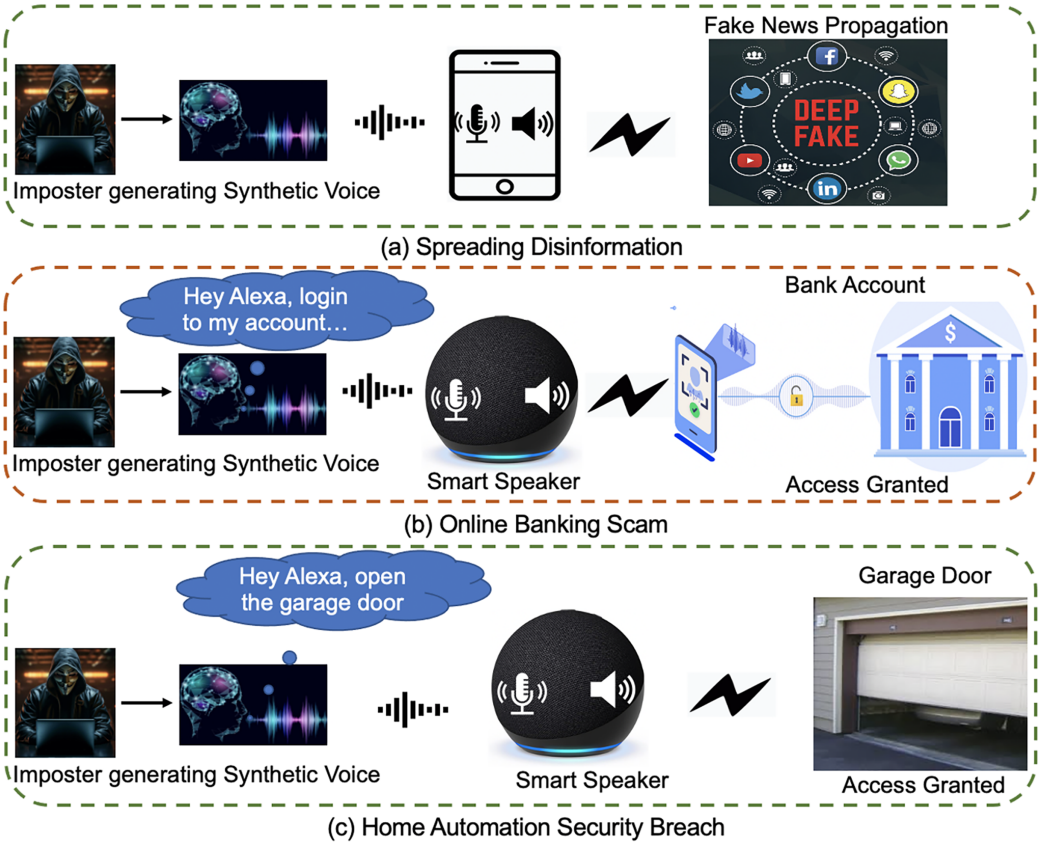

Figure 1 **A few scenarios of audio deepfakes attacks.** Full-size DOI: 10.7717/peerj-cs.3670/fig-1

emphasized the critical need to enhance the resilience of ASV systems against sophisticated logical attacks to safeguard financial transactions and prevent fraudulent activities. Furthermore, incidents involving voice synthesis and conversion have demonstrated the potential for deception in diverse contexts. For example, instances of manipulated audio impersonating public figures or political leaders underscore the broader societal implications of voice-based deception (*Kassis & Hengartner, 2021*). Developing ASV systems that can effectively distinguish between genuine and synthetic voices becomes crucial not only for security applications but also for ensuring the integrity of communication channels in various domains (*Khan et al., 2023b*; *Yu et al., 2023*).

Researchers have designed various techniques (*Nguyen-Vu et al., 2023*) for the effective recognition of bonafide and spoof voices. Initially, spectral feature analysis-oriented approaches were utilized for voice spoofing detection, involving the examination of frequency components within the voice signal (*Meriem, Messaoud & Bahia, 2023*). While this method is effective to some extent, it comes with inherent limitations. One key challenge is the vulnerability to advanced voice synthesis techniques that can precisely mimic natural spectral features. As attackers leverage sophisticated tools to generate artificially crafted voiceprints, distinguishing between genuine and synthetic voices based solely on spectral features becomes increasingly difficult (*Masood et al., 2023*). Moreover,

spectral analysis alone struggles to capture distinctive traits in manipulated voices, especially when attackers intentionally mimic the spectral characteristics of the target voice. The limitations in spectral feature analysis emphasize the need for more effective techniques to enhance the robustness of voice spoofing detectors. Machine Learning (ML) has proven to be a fundamental tool for effective voice spoofing detection (*Zhou et al., 2022*). ML techniques, particularly supervised learning models, were trained on datasets containing genuine and spoofed voice samples, enabling the system to learn patterns indicative of deception (*Iqbal et al., 2022*). These models offer a data-driven solution to voice spoofing detection by identifying the required patterns between authentic and manipulated voices; however, they lack scaling and generalization power. The limitations of traditional ML methods have led to a shift towards more sophisticated Deep Learning (DL) techniques. DL models can automatically extract complex features from the voice signal, enabling them to learn effective characteristics indicative of voice spoofing. Additionally, the ability of deep networks to adapt and generalize to diverse datasets enhances their effectiveness in handling evolving spoofing techniques (*Almutairi & Elgibreen, 2022*; *Cuccovillo et al., 2022*).

Current methods face several limitations in distinguishing between bonafide and spoofed audio, particularly in the aspect of increasingly sophisticated spoofing techniques. Between both categories of voice spoofing, detecting logical voice spoofing (synthesis and conversion) presents a unique set of challenges compared to recognizing physical voice spoofing. While physical voice spoofing involves the playback of pre-recorded audio, often exhibiting noticeable artifacts and microphone fingerprint traces, voice synthesis and conversion alter the characteristics of the person's voice dynamically without microphone fingerprint traces (*Liz-López et al., 2024*). In voice synthesis, entirely artificial voices are created, making it challenging to identify anomalies typically associated with replayed audio. Voice conversion, on the other hand, transforms one person's voice into another seamlessly, maintaining the natural flow and pace, thus becoming more challenging than voice synthesis for detectors (*Cohen et al., 2022*). The absence of evident cues or artifacts, coupled with the potential for highly convincing output, makes the detection of synthesized or converted voices inherently more complex. Attackers continually refine their approaches to mimic natural speech patterns with extraordinary accuracy. Traditional detection approaches, positioned around pattern recognition and anomaly detection, struggle to discriminate these dynamically altered voices, requiring advanced detection techniques to examine deeper into the semantic and contextual aspects of speech.

In this work, we designed a system to detect logical voice spoofing/deepfakes attacks by presenting a more reliable and effective strategy to overcome the above-listed challenges. This study seeks to address the following critical scientific questions: (1) How can we enhance the detection of logical access (LA) and deepfake (DF) attacks using advanced DL techniques? (2) What architectural modifications are most effective in improving the robustness and adaptability of RawNet2 for handling diverse spoofing scenarios? (3) Can the proposed model generalize effectively across datasets and attack types, ensuring consistent performance under varying codec and channel conditions? To answer these questions, we have introduced DeepRawNet, which addresses the limitations of the base

RawNet2 model. The novelty of our work lies in the strategic architectural redesign and collaborative interaction of Parametric Rectified Linear Unit (PReLU), transpose convolution layers, and LogSoftmax components, transforming RawNet2 from a generic countermeasure baseline into a more generalizable, fine-detail-preserving, and robust anti-spoofing network. The key contributions of our approach are:

- **Enhanced Feature Extraction and Model Adaptability:** The alterations introduced in the model, like increasing the negative slope in the Fixed Sinc filters, the employment of PReLU, and the substitution of the simple convolution layer with a transpose convolution layer in the residual block, together promote a more robust feature representation, resulting in preventing dead neurons and addressing downsampling issues.
- **Enhanced Stability:** The strategic improvements in our DeepRawNet's architecture contribute to improved stability and efficiency in computations by facilitating more stable and numerically efficient training and inference processes.
- **Enhanced VC Sample Detection:** Our method achieves superior performance over the base RawNet2 model and other contemporary methods in detecting VC samples within the ASVspoof2019-LA dataset.
- **Empowered Generalization:** The cumulative effect of these adaptations empowers our DeepRawNet model over the base RawNet2 with enhanced generalization ability for recognizing a variety of voice spoofing attacks on different corpora.
- **DeepRawNet-Verifier Tool:** We also propose a robust tool for the reliable detection of TTS and VC audio deepfakes from bonafide samples.

The remaining article is formatted as follows: the work from the history related to audio spoof detection is provided in 'Related Work', the proposed framework and experimental results are discussed in 'Materials and Methods'. A detailed discussion is provided in 'Discussion', while the conclusion is provided in 'Conclusions'.

RELATED WORK

This part holds the critical analysis of the existing audio spoofing detection approaches. These methods are broadly categorized as conventional ML and DL-based approaches.

Sinha, Dey & Saha (2024) discussed the growing threat of spoofing attacks targeting voice recognition systems and explored self-supervised learning (SSL) as a potential solution. Initially, SSL-based models show significant performance improvements over supervised learning (SL) baselines. However, comparative analysis reveals instances where SL outperforms SSL. To address this, the study proposed a fusion approach combining SL and SSL models, resulting in reduced Equal Error Rate (EER) and surpassing state-of-the-art (SOTA) frameworks. However, it may not generalize well across diverse attack types. *Chakravarty & Dua (2023)* suggested an ML approach that employed acoustic Ternary Pattern (ATP) and Local Binary Pattern (LBP) features in combination with various spectral features like MFCC, Constant Q Cepstral Coefficients (CQCC), Gammatone Cepstral Coefficients (GTCC), and Perceptual Linear Prediction (PLP). The fused features

were passed to the classification unit to perform the categorization task. The approach shows the best results for the ATP-LBP-GTCC along with the long short-term memory (LSTM) classifier. In *Meriem, Messaoud & Bahia (2023)*, Local Phase Quantization (LPQ), Heterogeneous Auto-Similarities of Characteristic (HASC), and Binarised Statistical Image Features (BSIF) were applied on audio spectrograms with the support vector machine (SVM) (*Chaabane et al., 2022*) to recognize the bonafide and spoofed audio samples. The approaches (*Meriem, Messaoud & Bahia, 2023*; *Chakravarty & Dua, 2023*) provide an efficient solution, but need performance enhancement. *Neelima & Prabha (2023)* proposed an ML method incorporating a textural feature descriptor for audio spoofing detection; however, it lacks generalization power. *Javed et al. (2022)* introduced an anti-spoofing framework that employed a textural feature descriptor, Acoustic-Ternary Co-occurrence Patterns (ATCoP) with GTCC. The method performed well in recognizing distortions and artifacts in cloned replays, including those resulting from compression in multi-hop attacks. The work performed well in recognizing spoof audio samples; however, the model needs improvement in the generalizability aspect. *Hamza et al. (2022)* focused on detecting deepfake audio through the utilization of MFCC features with machine learning algorithms. The approach shows the highest classification results with the SVM classifier; however, the model needs evaluation on a larger dataset. *Xue et al. (2022)* presented a method that combined F0 information (fundamental frequency) with features extracted from the real and imaginary components of the spectrogram. The approach aimed to leverage both pitch-related information and complex spectral features for a more robust audio deepfake detection system; however, classification results need improvement.

Existing countermeasures have also employed DL-based approaches for audio spoofing detection. *Liu, Zhang & Gao (2024)* employed a DL framework introducing multi-space channel representation learning. The work performs well in recognizing the bonafide audio samples from the manipulated, however, the approach lacks explainability power. In *Liu et al. (2024)*, the introduction of the Generalizing Speaker Verification Systems (G-SASV) job against spoofing attacks was defined, covering baseline systems, decision policies, and evaluation metrics. The approach performs well for audio spoofing detection; however, performance needs further improvements. *Dişken (2024)* suggested a DL approach named SE-Res2Net model with log power spectrogram features to classify the bonafide and spoofed audio samples. This model achieves better results for the ASVspoof2019 LA set; however, the approach lacks generalization power. *Chakravarty & Dua (2024)* employed an enhanced ResNet50 on audio Mel spectrograms, followed by minimizing the keypoints space by employing Linear Discriminant Analysis (LDA), which was then passed to train various ML classifiers, like SVM, RF, K-Nearest Neighbor (KNN), and Naive Bayes (NB). The approach was evaluated on ASVspoof2021 deepfake partition and showed effective results for audio spoofing detection; however, unable to tackle diverse spoofing attacks. *Boyd, Fahim & Olukoya (2023)* employed numerous DL models like Convolutional Neural Networks (CNNs) (*Alsalhi & Almehmadi, 2024*), WaveNet, and Recurrent Neural Network (RNN) for dense keypoints calculation and performing the classification. The approach needs extensive samples for model tuning. Another DL approach (*Doan, Hong & Jung, 2023*) employed a Generative Adversarial Network (GAN)-

based vocoder framework for computing the representative set of sample features, which were then recognized as bonafide or spoofed. *Ma et al. (2023)* employed a ConvNeXt DL approach that processed raw audio and exhibited better generalization ability. The approaches (*Doan, Hong & Jung, 2023*; *Ma et al., 2023*) were tested on the ASVspoof2019-LA data sample; however, the results need further improvements. A robust DL-based voice spoofing detection system employing Center Lop-Sided Local Binary Patterns (CLS-LBP) and LSTM (*Abdelhamid et al., 2022*) was proposed in *Dawood et al. (2022)* to discern genuine voices from spoofed ones, but it lacks the generalization power. *Huang et al. (2025)* introduced a spoofing detection method implementing latent space refinement and augmentation to improve the generalization aptitude. In *Zhang et al. (2023)*, a pre-trained wav2vec2 model was used for dense keypoints computation, and keypoints merging unit was applied for back-end categorization. *Xue & Zhou (2023)* employed a densely connected CNN with a squeeze and excitation block (SE-DenseNet), with a keypoints merging mechanism. In *Huang & Pun (2024)*, a hybrid approach based on computing deep and Mel spectrogram features through parallel paths of CNNs and short-time Fourier transform (STFT) was presented. The features were then combined through a max-pooling layer, along with an attention technique to emphasize crucial aspects of samples. The approaches (*Zhang et al., 2023*; *Xue & Zhou, 2023*; *Huang & Pun, 2024*) enhance the audio spoofing classification results; however, with an increased computational cost. DeepLASD framework based on RawNet was introduced in *Al-Tairi et al. (2025)* for spoofing detection tasks. *Dua et al. (2025)* introduced an approach combining Mel spectrogram and Cochleagram features with ResNet and InceptionV3. The performance of approaches (*Al-Tairi et al., 2025*; *Dua et al., 2025*) needs to be evaluated for the diverse and unknown spoofing attacks. Many researchers have explored Audio Anti-Spoofing using Integrated Spectro-Temporal (AASIST) Graph networks (*Jung et al., 2022*), including SSL-AASIST (*Ge et al., 2024*), AASIST2 (*Zhang et al., 2024*), and AASIST3 (*Borodin et al., 2024*), SSL and spectrogram feature fusion with AASIST (*Rishith Sadashiv et al., 2025*), for audio spoofing detection. AASIST leverages integrated spectro-temporal representations to effectively capture the relevant cues that distinguish bonafide and spoofed audio, focusing on joint feature extraction and classification. SSL-AASIST extends this capability by incorporating self-supervised learning, which enhances performance in scenarios with limited labeled data by leveraging large amounts of unlabeled audio. AASIST2 and AASIST3 further refine the architecture with advanced optimization techniques and additional features to improve generalization across diverse datasets. However, these models face limitations, such as computational complexity, reliance on high-quality feature extraction, susceptibility to novel spoofing techniques not represented in training data, and lower performance in low-resource or noisy environments, necessitating further innovations to address these challenges effectively.

After providing a detailed analysis of the existing approaches, it is evident that early audio spoofing detection relied on traditional ML with handcrafted features, but showed limited adaptability and generalizability. Subsequently, the field experienced a transformative shift with the application of DL techniques, leading to significant improvements through automatic feature learning. However, DL-based countermeasures

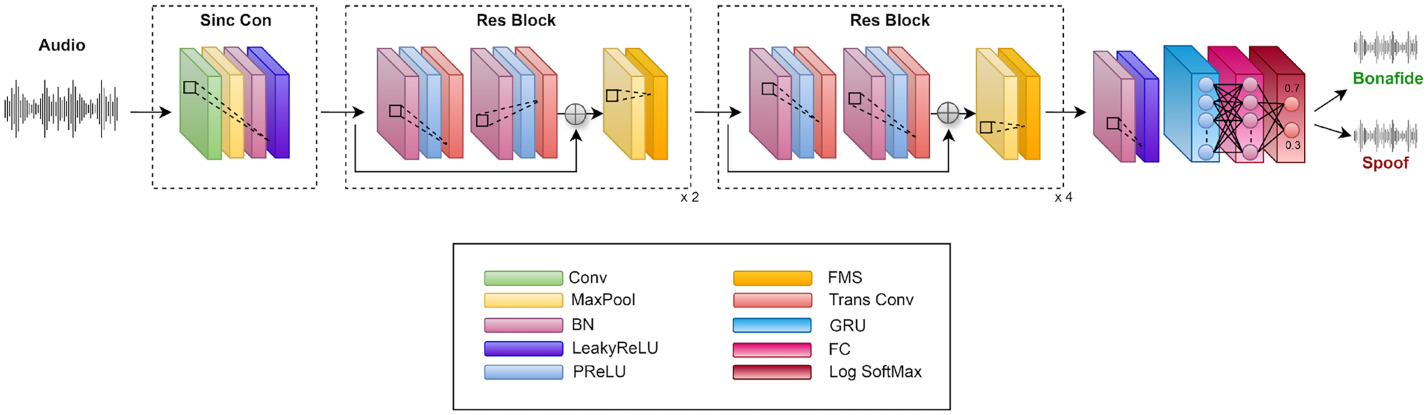

Figure 2 **Flow of the presented model.** Full-size DOI: 10.7717/peerj-cs.3670/fig-2

are still facing challenges due to the increased realism of the generated vocal samples. These limitations present a need to develop a more effective approach to address evolving audio spoofing attacks.

MATERIALS AND METHODS

We introduced an end-to-end DL framework, DeepRawNet (Fig. 2), to enhance the audio spoofing detection performance over the ASVspoof base RawNet2 model. Specifically, the baseline RawNet2 (*RawNet2 Code, 2021*) is systematically improved through architectural changes to enhance its discriminative capability against sophisticated spoofing attacks. The negative slope in the Fixed Sinc filters is increased to prevent dead neurons, which provides a more dynamic and effective representation of features across the entire input space. We propose to use the PReLU activation in the residual blocks over the Leaky ReLU in RawNet2 to present a learnable negative slope to mitigate the risk of dead neurons and ensure more effective feature extraction. Furthermore, a pivotal modification involves substituting the simple convolution layer with a transpose convolution layer in the residual block to address downsampling issues and to better preserve fine-grained temporal information vital for capturing complicated patterns in raw audio. We employ the LogSoftmax activation in the final layer of the network to transform the raw model outputs into logarithmic probabilities, facilitating more stable and numerically efficient computations during training and inference. Together, these modifications empower the proposed DeepRawNet with improved discriminative power, adaptability, robust learning capabilities, generalization ability, and enhanced capacity to learn sequential dependencies within audio, making it a more effective solution for audio spoofing detection.

RawNet2

RawNet2 (*Tak et al., 2021*) is an extended form of RawNet (*Jung et al., 2019*), which is a well-known neural network architecture designed for processing raw audio waveforms directly, without the need for traditional feature extraction methods. RawNet aims to capture complex temporal patterns found in audio signals, providing a more direct and detailed representation for tasks like speech processing and speaker recognition. This

Table 1 Architectural comparison of RawNet and RawNet2.

RawNet			RawNet2		
Layer	**Input**	**Output**	**Layer**	**Input**	**Output**
Stride Con	Conv(3,3,128) BN LeakyReLU	(19,683, 128)	Sinc Con	Sinc(251,1,128) MaxPool(3) BN LeakyReLU	19,683, 128
Res block	{Conv(3,1,128) BN LeakyReLU Conv (3, 1, 128) BN LeakyReLU MaxPool(3)} ×2	(2,187, 128)	Res block	{BN LeakyReLU Conv(3,1,128) BN LeakyReLU Conv(3,1,128) ———————————————— MaxPool(3) FMS}×2	(2,187, 128)
Res block	{Conv(3,1,256) BN LeakyReLU Conv(3,1,156) BN LeakyReLU MaxPool(3)} ×4	(27, 256)	Res block	{BN LeakyReLU Conv(3,1,256) BN LeakyReLU Conv(3,1,256) ———————————————— MaxPool(3) FMS}×4	(27, 256)
GRU	1,024	(1,024)	GRU	1,024	(1,024)
Speaker embedding	128	(128)	Speaker embedding	1,024	(1,024)

architecture typically consists of convolutional layers for feature extraction, recurrent layers (such as LSTM) for capturing temporal dependencies, and fully connected layers for final predictions. The direct use of raw waveforms allows RawNet to learn hierarchical features from the audio and exhibit improved performance for anti-spoofing. However, the original RawNet faces issues with computational demands and overfitting due to processing raw waveforms, which requires the computation of a more relevant set of sample features. RawNet2 aims to overcome limitations inherent in the original RawNet model, particularly addressing challenges related to computational efficiency, generalization, and model parameters. RawNet2 introduces Filter-wise Feature Map Scaling (FMS) in residual blocks, allowing dynamic adjustment of feature map sizes. This integration reduces model parameters, enhancing computational efficiency and potentially improving generalization across diverse datasets. Additionally, the multiplicative FMS in RawNet2, resembling an attention mechanism, suggests an innovation in capturing relevant information, which contributes to addressing limitations related to interpretability and the effectiveness of attention mechanisms that were present in the original RawNet. A structural comparison of RawNet and RawNet2 is provided in Table 1. RawNet2 architecture comprises several key components, including residual blocks, a gated recurrent unit (GRU), a fully connected layer, and an output layer. The initial step involves

the extraction of the input feature sequence from the frame-level representation through the utilization of residual blocks. Subsequently, the GRU is employed to aggregate the frame-level representation into an utterance-level representation, allowing for comprehensive sequence analysis and discrimination. The resultant representation is then passed through a fully connected layer. During the evaluation phase, a Softmax activation function is applied after the fully connected layer to classify bonafide or spoofed audios.

Proposed DeepRawNet method

The primary limitation of the baseline RawNet2 model is its restricted generalization and robustness when encountering unseen spoofing attacks. Moreover, another limitation is its inability to fully extract deeper features from examined audio, which impacts the model's ability to learn complicated patterns of deepfake speech. To overcome these issues, our work proposes architectural improvements rather than a simple combination of existing methods. Each modification has a specific role in improving model performance. PReLU activation introduces a learnable negative slope, effectively preventing neuron inactivation and improving adaptability to diverse spoofing patterns. The transpose convolution layers facilitate learned upsampling, allowing better reconstruction of subtle spoofing artifacts and improving generalization to unseen attack types. The Log-Softmax function enhances numerical stability and provides well-calibrated output probabilities, improving decision reliability. Thus, the proposed DeepRawNet framework specifically addresses the limitations of RawNet2 through an effective architectural redesign, resulting in improved robustness, generalization, and discriminative performance. This demonstrates a research-driven contribution rather than mere development work. A pictorial representation showing the comparison of the RawNet2 and proposed DeepRawNet approach is shown in Fig. 3.

In the first step, we increased the negative slope value of LeakyReLU in the Fixed Sinc filters part. We experimented with various values of the negative slope parameter, and after testing values ranging from 0.1 to 0.6, we determined that a slope value of 0.5 attained the most optimal results. Increasing the negative slope in Leaky ReLU is a significant step, as it addresses one of the limitations inherent in traditional ReLU activation. The standard ReLU function sets all negative values to zero, leading to dead neurons and potential information loss during training. By introducing a small, non-zero negative slope in Leaky ReLU, typically denoted by a parameter (commonly denoted as alpha), the activation function allows a small gradient for negative inputs. This modification is crucial as it mitigates the issue of dead neurons and ensures that all neurons contribute to the learning process. It prevents neurons from becoming entirely inactive, facilitating the flow of gradients during backpropagation and enabling the model to learn from all data points, including those with negative input values. This helps in capturing more diverse features and representations, enhancing the model's expressiveness. Moreover, the increased negative slope addresses the vanishing gradient problem to some extent, making it particularly beneficial in deep neural networks (DNNs). In our scenario, where a significant portion of the input values can be negative, the increased negative slope allows for more effective learning, improving the overall training dynamics.

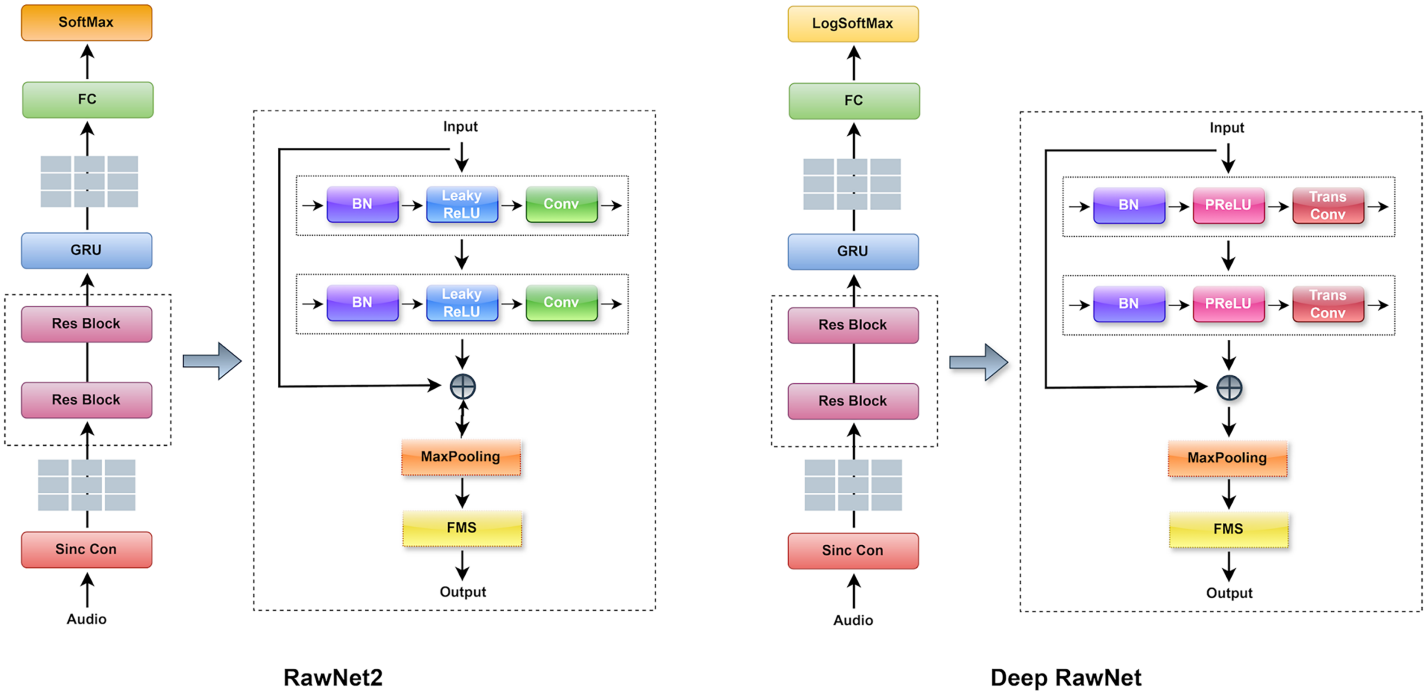

Figure 3 **Comparison of the RawNet2 and proposed DeepRawNet approach.** Full-size DOI: 10.7717/peerj-cs.3670/fig-3

Next, we introduced two changes within the residual block of the RawNet2. The residual block in RawNet2 serves as a foundational component in the network architecture, crucial for processing raw audio waveforms effectively. Comprising a skip connection and a main processing path, the residual block enables the extraction of complex features by combining information from both paths. This design is pivotal in overcoming the vanishing gradient problem, a common challenge in training DNNs. The skip connection facilitates the smooth flow of gradients, allowing for the effective training of deeper architectures without degradation issues. However, the residual block is still facing some issues, like when using LeakyReLU in the original residual block of RawNet2, one potential issue is the possibility of dead neurons. In LeakyReLU, the negative slope is fixed, typically at a small constant value. This means that for all negative inputs, the gradient is scaled by a constant factor, preventing complete neuron inactivity as in traditional ReLU, but still potentially leading to suboptimal learning dynamics. The fixed negative slope in LeakyReLU may not be ideal for capturing the effective patterns present in raw audio data. Some neurons remain relatively inactive throughout training, especially if the data distribution requires a more flexible adjustment of the negative slope (*Mastromichalakis, 2020*). Consequently, this results in a partial loss of representational capacity, hindering the model's ability to learn discriminative features from the raw audio waveforms effectively. To address this issue, we introduced a PReLU with a learnable parameter for the negative slope in our proposed DeepRawNet method. PReLU's adaptability allows the model to learn the optimal negative slope during training, preventing the occurrence of dead neurons and ensuring a more dynamic and effective utilization of neurons across the entire

input range (*QingJie & WenBin, 2017*). This adjustment contributes to improved learning capabilities, especially in scenarios where a fixed slope is suboptimal for capturing the complexity of raw audio features.

Secondly, we incorporated a transpose convolution layer in the residual block. The replacement of a simple convolution layer with a transpose convolution layer in the residual block of RawNet2 addresses issues related to downsampling and the loss of fine-grained temporal information. The use of standard convolutional layers typically involves operations like pooling, which downsamples the input and may lead to a reduction in the temporal resolution of the features. One significant issue with downsampling operations is the potential loss of fine temporal details, especially crucial in tasks involving sequential data, such as speech processing (*Stergiou, Poppe & Kalliatakis, 2021*). In the context of raw audio waveforms, preserving the fine-grained temporal information is essential for capturing effective patterns and temporal dependencies. By incorporating transpose convolution layers, also known as deconvolution or fractionally stridden convolution layers, our DeepRawNet aims to upsample the features, counteracting the downsampling effects introduced by standard convolution layers. Transpose convolution layers help in the reconstruction of higher-resolution feature maps, allowing the model to retain more detailed temporal information and better capture the sequential dependencies present in raw audio signals.

A GRU layer comprising 1,024 hidden nodes is utilized to consolidate frame-level representations into a unified utterance-level representation. Unlike RawNet 2, which generates embeddings directly, the GRU output is followed by an additional fully connected layer before reaching the output layer. Subsequently, a LogSoftmax activation function is employed to yield two-class predictions, distinguishing between bonafide and spoof in place of Softmax. In contrast to Softmax, LogSoftmax stands out as a particularly advantageous choice for tasks in raw audio processing within deep learning applications (*Wei, Yao & Meng, 2023*). The key distinction lies in LogSoftmax's ability to enhance numerical stability by working within the logarithmic space. This proves crucial in mitigating challenges associated with numerical precision during probability calculations, especially when dealing with raw audio data. Unlike Softmax, LogSoftmax not only streamlines computational efficiency by consolidating exponentiation and normalization but also simplifies gradient computations during backpropagation, offering a more efficient processing approach. The LogSoftmax output, interpreted as log probabilities, provides a better understanding of the model's confidence in predictions, a feature particularly valuable in the complexities of raw audio processing. Additionally, LogSoftmax plays a pivotal role in reducing the risk of numerical underflow or overflow, contributing to a more steadfast and reliable probability estimation process compared to Softmax.

Dataset

For model evaluation, we have utilized the standard benchmark datasets, *i.e.*, ASVspoof2019 (https://datashare.ed.ac.uk/handle/10283/3336) (*Borodin et al., 2024*), and ASVspoof2021 (https://www.asvspoof.org/index2021.html) (*Tak et al., 2021*) in audio

Table 2 Overview of datasets.

Datasets	Training			Development			Evaluation		
	Bonafide	Spoof	Total	Bonafide	Spoof	Total	Bonafide	Spoof	Total
ASV2019-LA	2,580	22,800	25,380	2,548	22,438	24,986	7,355	64,578	71,933
ASV2021-LA	—			—			14,816	133,360	148,176
ASV2021-DF	—			—			14,869	519,059	533,928

spoofing/deepfakes detection. ASVspoof2019 serves as a comprehensive benchmark dataset developed to evaluate the efficacy of ASV systems when opposed to various spoofing attacks. Comprising a diverse collection of speech recordings, the dataset encompasses both genuine utterances and a spectrum of spoofing techniques, including replayed speech, voice conversion, and speech synthesis. Through a balanced distribution of genuine and spoofed samples across different attack scenarios, ASVspoof2019 ensures a thorough evaluation of ASV systems under realistic conditions. Public availability of ASVspoof2019 facilitates collaborative efforts within the research community, enabling the development of more robust ASV technologies capable of effectively detecting and mitigating the risks associated with spoofing attacks in real-world scenarios. For our model evaluation, we have utilized the ASVspoof2019 logical access (ASV2019-LA) dataset, which is specifically designed for evaluating ASV systems in logical access scenarios. It consists of training, development, and evaluation sets, each comprising 25,380, 24,986, and 71,933 audio samples. ASVspoof2021 is also a benchmark dataset designed for evaluating the performance of ASV systems against spoofing attacks. It represents an evolution from previous editions, featuring a more diverse and challenging set of spoofing attacks, including replay, voice conversion, and synthesis techniques. The dataset provides a comprehensive evaluation platform with standardized protocols, facilitating fair comparison and benchmarking of ASV technologies. In our case, we have utilized the samples of ASV2021-LA (logical access) and ASV2021-DF (deepfake). ASV2021-LA consists of 148,176 samples, whereas ASV2021-DF comprises 533,928 audios. ASV2021-LA and ASV2021-DF are subsets of the ASVspoof2021 dataset designed for specific evaluation scenarios. ASV2021-LA focuses on spoofing attacks relevant to logical access with codec and transmission channel variability, unlike ASVspoof2019-LA, where these variations do not exist. It offers a balanced distribution of genuine and spoofed speech samples across various conditions, enabling rigorous evaluation of ASV systems' robustness in logical access settings. ASV2021-DF, on the other hand, emphasizes deepfake attacks where speech is artificially generated using advanced synthesis techniques and contains compressed audios. This subset challenges ASV systems to detect increasingly sophisticated spoofing attempts, contributing to advancements in anti-spoofing technologies. The overview of datasets in terms of the number of utterances is provided in Table 2.

Evaluation metrics

We assessed the efficacy of various audio deepfake detection systems using two key evaluation metrics: minimum Tandem Detection Cost Function (min t-DCF) and Equal

Table 3 Performance evaluation of proposed DeepRawNet.

Test dataset	min-tDCF	EER (%)
ASV19-LA–Eval set	0.1275	2.80
ASV21-LA–Eval set	0.3666	7.77
ASV21-DF–Eval set	—	20.72

Error Rate (EER). Min-tDCF offers a comprehensive evaluation by providing the assessment of a tandem system while keeping the countermeasure (CM) and ASV systems isolated from each other (*Jung et al., 2019*). Additionally, EER represents the point where the false rejection rate (FRR) and false acceptance rate (FAR) are equal, providing a balanced measure of performance. We also utilized accuracy to show the performance of our method in some of the experiments. Accuracy represents accurately classified audio samples over the total audio samples, thus indicating our method's ability to accurately distinguish between bonafide and spoofed audio samples.

Experimental settings

The proposed model is implemented in the PyTorch framework. Experiments were conducted on a machine with 32 GB RAM, a 12-core processor (3.70 GHz), and an NVIDIA GeForce RTX 3090 GPU running Windows 11. The network undergoes training with an ADAM optimizer using a learning rate of 0.0001 and weight decay of 0.0001, employing cross-entropy loss and a batch size of 32. The proposed model directly processed the raw audio signals. No additional data preprocessing was applied. The model is trained for 100 epochs using the train and development sets, where the train set is utilized as training data and the development set is used as validation data. However, the model weights saved at the last epoch are utilized to evaluate it on the evaluation set.

Performance evaluation

To evaluate the performance of our method, we performed two different experiments utilizing ASVspoof2019 and ASVspoof2021 datasets. The model is trained utilizing the ASV2019-LA dataset, while for testing, we have used the ASV2019-LA, ASV2021-LA, and ASV2021-DF datasets. By evaluating the model's performance on the ASV2021-LA and ASV2021-DF datasets, which are distinct from the training dataset (ASV2019-LA), we aim to understand how well the model performs on unseen data and its ability to detect spoofing attacks in different contexts, thus assisting in assessing the generalization capability and cross-dataset performance. The results reported in Table 3 show that our approach attained effective results over both datasets. For the ASV2019-LA dataset, we have attained min-tDCF and EER of 0.1275% and 2.80%. These results highlight the effective audio-spoofing detection performance of the proposed DeepRawNet. For the cross-dataset ASV2021-LA, our model achieved min-tDCF and EER of 0.3666% and 7.77%, while for the ASV2021-DF, our approach acquired an EER of 20.72%. It is quite evident from the results reported on ASV2021 datasets that the proposed approach shows effective generalization power while detecting the unknown logical access and deepfake

Table 4 Results on specific unseen attacks in evaluation set of ASVspoof2019-LA dataset.

Attacks	Accuracy (%)	EER (%)	min-tDCF
A07	98.95	0.22	0.0538
A08	97.11	3.24	0.1232
A09	98.95	0.11	0.1831
A10	98.81	0.42	0.0609
A11	98.90	0.38	0.0588
A12	98.92	0.42	0.0614
A13	98.95	0.11	0.0519
A14	98.95	0.23	0.0536
A15	98.94	0.27	0.0560
A16	98.86	0.63	0.0645
A17	89.04	7.33	0.7333
A18	72.20	16.46	0.8376
A19	98.59	1.37	0.1295

attacks. Moreover, it has been found from the reported values that we attained better performance for the LA samples because of the complexity of the DF samples, which involves compression attacks resulting in higher EER than the LA sets.

Algorithm-wise model evaluation–ASV2019-LA–eval

This experiment is conducted to evaluate the proposed model's ability to detect and classify spoofed audios generated by specific attacks within the ASV19-LA dataset. ASVspoof2019-LA eval set contains the spoofed audios generated using different unknown attacks (A07 to A19). This algorithm-wise evaluation provides insights into how effectively the model can generalize and detect spoofing attacks it has not encountered during training, thus contributing to a comprehensive understanding of its performance in real-world scenarios. Three evaluation metrics, namely EER, min-tDCF, and accuracy, are computed, and the attained results are provided in Table 4. Our evaluation across various unknown attack techniques (A07 to A19) in the ASV19-LA dataset yielded promising results, showcasing the effectiveness of our approach in detecting a wide range of spoofing techniques. A little performance degradation has been witnessed for A17 and A18 attacks, as these contain VC samples, which are difficult to locate due to their high realism. The results in Table 4 highlight the robustness of our model, demonstrating its ability to accurately identify spoofed audio samples generated using different attack algorithms.

Comparison with baseline models

This section presents the overall performance comparison, VC, and TTS attacks comparison utilizing ASV2019-LA, ASV2021-LA, and ASV2021-DF datasets.

Overall performance comparison

To assess the performance of the proposed DeepRawNet against the existing ASVspoof baseline models, *i.e.*, RawNet2, LFCC-LCNN, LFCC-GMM, and CQCC-GMM, we

Table 5 Comparison with baseline models utilizing the ASVspoof2021-DF dataset.

Model	Test dataset	EER (%)
Baseline–LFCC-LCNN	ASV2021–DF–Eval set	23.48
Baseline–LFCC-GMM		25.25
Baseline–CQCC-GMM		25.56
Baseline–RawNet2		22.38
DeepRawNet (Proposed)		20.72

performed three comparison experiments utilizing the ASVspoof2019-LA, ASVspoof2021-LA, and ASVspoof2021-DF datasets. Each baseline model represents a distinct approach to spoofing detection, ranging from traditional feature-based methods like LFCC and CQCC with GMM to more recent DL-based architectures like the conventional RawNet2. By comparing our model's performance against these established baselines, we aimed to provide a comprehensive assessment of its efficacy in detecting bonafide and spoofed audio.

In the first experiment, we conducted a critical comparison of the performance of our model with baseline models, specifically on the ASV2021-DF Eval set. To accomplish this, we compared the proposed method with several baseline models, namely RawNet2, LFCC-LCNN, LFCC-GMM, and CQCC-GMM. The scores in Table 5 clearly show that our proposed approach attained the lowest EER of 20.72% over the ASVspoof2021-DF dataset, indicating the effectiveness of our approach while detecting unknown audio deepfake samples.

In the next experiment, we compared our method with the SOTA baseline approaches on the ASVspoof2021-LA dataset, and provided scores of min-tDCF and EER in Table 6. Both min-tDCF and EER scores show that our approach outperforms the baseline approaches. Specifically, the baseline LFCC-GMM approach attained the lowest results, whereas the LFCC-LCNN approach performed second-best. In contrast, our model shows a significant improvement in detecting unseen logical access attacks with the min-tDCF and EER scores of 0.3666 and 7.77%, which clearly shows the robustness of our approach.

The results reported in Tables 5 and 6 show that our approach performs well for unseen LA and DF samples from the ASVspoof2021 dataset. The LA and DF tasks from the ASVspoof2021 dataset contain bonafide and spoofed utterances generated using TTS and VC algorithms, where the LA set contains samples with coding and transmission variations, while the DF set contains compressed samples. The better performance of our approach over both sets indicates that our approach is robust and generalized enough to handle the compression, codec, and transmission channel variability.

In the third experiment, we compared our method with the baseline approaches on the ASVspoof2019-LA dataset, and the acquired analysis is provided in Table 7. For this experiment, our approach outperforms the comparative approaches by attaining the minimum EER of 2.80%. However, the proposed DeepRawNet attains the second-best min-tDCF on the ASV2019-LA dataset, compared to the baseline methods.

Table 6 Comparison with baseline models utilizing the ASVspoof2021-LA dataset.

Model	Test dataset	min-tDCF	EER (%)
Baseline–LFCC-LCNN	ASV2021–LA–Eval set	0.3445	9.26
Baseline–LFCC-GMM		0.5758	19.30
Baseline–CQCC-GMM		0.4974	15.62
Baseline–RawNet2		0.4257	9.50
DeepRawNet (Proposed)		0.3666	7.77

Table 7 Comparison with baseline models utilizing ASVspoof2019-LA dataset.

Model	Test dataset	min-tDCF	EER (%)
Baseline–LFCC-LCNN	ASV2019–LA–Eval set	0.100	5.06
Baseline–LFCC-GMM		0.212	8.09
Baseline–CQCC-GMM		0.237	9.57
Baseline–RawNet2		0.129	4.66
DeepRawNet (Proposed)		0.1275	2.80

All comparisons provided in Tables 5–7 indicate that our approach yields significant performance improvements over the base models on all datasets. The baseline models exhibit various limitations that hinder their effectiveness in deepfake audio detection. Traditional methods, such as LFCC-LCNN and LFCC-GMM, rely on handcrafted features that are unable to capture the complex patterns present in deepfake speech. Similarly, the CQCC-GMM approach faces challenges in representing the diverse acoustic properties of deepfake audio. These feature-based methods often struggle with generalization across different datasets and lack the adaptability to learn complex representations directly from raw audio. On the other hand, while conventional RawNet2 represents an advancement by leveraging deep learning techniques, it still encounters issues such as vanishing gradients or dead neurons, leading to suboptimal performance and limited generalization. Further, conventional RawNet2 is not proficient in fully extracting deeper features from fake audio, and it lacks capturing intricate characteristics embedded in deceptive audio, which impacts the model's capacity to distinguish subtle distinctions indicative of deepfake speech (*Li et al., 2023*). Our proposed approach addresses these limitations by introducing a novel architecture with improved discriminative feature learning, fine-grained temporal information preservation, and enhanced robustness against diverse spoofing attacks, as validated through comparison experiments.

Performance comparison in detecting VC-based attacks

In this experiment, we specifically focus on VC samples, a challenging subset of the ASV19-LA and ASV21-LA datasets. VC samples pose a challenge for detection due to their close resemblance to genuine speech, evolving sophistication in synthesis techniques, and retention of original speaker characteristics. The objective is to assess the model's performance in detecting manipulated or converted voices, which are inherently difficult to detect due to their realistic nature. The performance of DeepRawNet for detecting VC

Table 8 Comparison with baseline models on VC samples (ASVspoof2019-LA).

Method	min-tDCF	EER (%)
LFCC-LCNN	0.5231	15.94
LFCC-GMM	0.6363	14.48
CQCC-GMM	0.7221	24.03
Baseline RawNet-2	0.6638	18.34
DeepRawNet (Proposed)	0.4905	13.46

Table 9 Comparison with baselines on VC samples (ASVspoof 2021-LA).

Model	min-tDCF	EER (%)
LFCC-LCNN	0.8483	23.02
LFCC-GMM	0.8608	22.15
CQCC-GMM	0.9444	32.15
Baseline RawNet2	0.8876	19.71
DeepRawNet (Proposed)	0.7308	14.95

attacks is compared against the baseline approaches, namely LFCC-LCNN, LFCC-GMM, CQCC-GMM, and RawNet2.

For the VC samples of the ASVspoof2019-LA dataset, the attained comparison is given in Table 8. Notably, our approach demonstrates significant improvement compared to all baselines, highlighting the robustness of our approach to the challenging voice conversion attacks in the domain of deepfake detection. Next, the performance of the approach was checked for the VC samples from the ASVspoof2021-LA dataset, and the obtained comparison in terms of EER and min-tDCF is reported in Table 9. The scores declare that our approach attained the best results over the contemporary approaches, thus signifying that our approach is robust to better tackle the codec variations induced in the ASV21-LA *corpus*. The results reported in Tables 8 and 9 highlight the effectiveness of our method in enhancing the model's ability to accurately detect voice conversion samples, thereby contributing to the advancement of deepfake audio detection technology.

Performance comparison in detecting TTS-based attacks

In this experiment, we checked the performance of our approach in detecting the TTS samples from the ASVspoof2019-LA and ASVspoof2021-LA datasets with several baseline approaches. TTS samples are particularly challenging due to the advanced synthesis techniques used to generate them, which can closely mimic the acoustic properties of genuine human speech while embedding subtle artifacts that are difficult to detect. Evaluating our model against such samples is crucial to ensure its reliability in real-world scenarios where sophisticated TTS systems are employed to create convincing fake audio.

For the TTS samples of the ASVspoof2021-LA dataset, the comparison results are reported in Table 10. Our results demonstrated a noticeable improvement in both EER and min-tDCF scores, showcasing the superior performance of our DeepRawNet in distinguishing between genuine and spoofed TTS audio samples. This highlights that the

Table 10 Comparison with baselines on TTS samples (ASVspoof2021-LA).

Model	EER (%)	min-tDCF
LFCC-LCNN	1.76	0.2014
LFCC-GMM	19.71	0.5384
CQCC-GMM	9.98	0.3826
Baseline RawNet2	6.08	0.3230
DeepRawNet (Proposed)	0.76	0.1826

Table 11 Comparison with baselines on TTS samples (ASVspoof 2019-LA).

Model	min-tDCF	EER (%)
LFCC-LCNN	0.0738	1.128
LFCC-GMM	0.0765	2.816
CQCC-GMM	0.3016	10.061
Baseline RawNet-2	0.0686	0.55
DeepRawNet (Proposed)	0.0722	0.64

proposed approach can capture intricate patterns within the compressed audio data that are indicative of TTS spoofing attempts.

In the subsequent experiment, we compared our TTS detection results on the ASVspoof2019-LA dataset with the baseline methods, and the comparison is presented in Table 11. Our method attained comparable results with the RawNet2, whereas it achieved better performance compared to other baseline methods, including LFCC-LCNN, LFCC-GMM, and CQCC-GMM, for the TTS attacks. Even though our DeepRawNet model performed slightly lower than the baseline RawNet2, however, the results are still satisfactory. This slight performance gap emphasizes the need for further fine-tuning and possibly incorporating additional techniques, such as advanced feature engineering or ensemble learning, to further improve the robustness of our method for TTS attacks.

Algorithm-wise model evaluation with the baseline models

This experiment is conducted to compare the algorithm-wise performance of the proposed DeepRawNet model on the ASVspoof2019-LA and ASVspoof2021-LA datasets with the baseline approaches (LFCC-LCNN, LFCC-GMM, CQCC-GMM, and RawNet2). This experiment is essential for validating the effectiveness of our DeepRawNet to detect deepfake audio across various attack techniques. For the ASVspoof2019-LA dataset, the attained comparison is provided in Table 12. Our method significantly improved the performance on the most challenging A17 unknown VC attack, along with the superior performance in detecting other challenging VC attacks, including A18 and A19. The baseline LFCC-LCNN performed best on most of the TTS attacks; however, our model achieved outstanding results for VC attacks. Moreover, the proposed DeepRawNet attained the lowest average EER across all attacks, compared to the baseline methods; however, the average min-tDCF is comparable to that of LFCC-LCNN. Achieving better performance metrics (specifically for VC attacks) compared to the baseline methods

Table 12 Algorithm-wise comparison of the DeepRawNet approach with the base models over the ASVspoof2019-LA dataset.

Attacks	DeepRawNet (ours)		RawNet-2 (Baseline)		LFCC-LCNN (Baseline)		LFCC-GMM (Baseline)		CQCC-GMM (Baseline)	
	min-tDCF	EER	min-tDCF	EER	min-tDCF	EER	min-tDCF	EER	min-tDCF	EER
A07	0.0538	0.22	0.0551	0.27	0.0237	0.71	0.0105	0.57	0.4980	15.44
A08	0.1232	3.24	0.1231	3.05	0.0787	2.62	0.1654	4.07	0.0551	1.59
A09	0.1831	0.11	0.1831	0.12	0.0300	0.11	0.0386	0.16	0.0020	0.02
A10	0.0609	0.42	0.0634	0.53	0.0758	0.77	0.2099	6.75	0.5680	17.27
A11	0.0588	0.38	0.0588	0.34	0.0219	0.65	0.0320	0.31	0.1935	5.71
A12	0.0614	0.42	0.0629	0.51	0.0144	0.45	0.0990	3.57	0.0749	2.04
A13	0.0519	0.11	0.0596	0.41	0.0194	0.63	0.0904	3.11	0.1291	3.97
A14	0.0536	0.23	0.0536	0.23	0.0124	0.37	0.1210	4.47	0.2360	7.12
A15	0.0560	0.27	0.0578	0.37	0.0958	0.38	0.1045	3.74	0.3156	9.89
A16	0.0645	0.63	0.0700	0.90	0.0256	0.88	0. 0810	0.42	0.1802	7.53
A17	0.7333	7.33	0.8321	13.15	0.9074	17.44	0.8071	11.66	0.8546	18.82
A18	0.8376	16.46	0.8929	19.59	0.9365	18.48	0.9287	17.97	0.9034	22.61
A19	0.1295	1.37	0.1767	2.65	0.1618	5.84	0.2896	9.19	0.8902	29.65
Average	0.1898	2.40	0.2069	3.24	0.1849	3.79	0.2291	5.08	0.3770	10.90

demonstrates the significance of our approach in enhancing the raw audio processing capabilities for spoofed audio detection.

Next, the results are reported for the ASVspoof2021-LA dataset, and the comparison is given in Table 13. For the ASVspoof2021-LA dataset, the baseline LFCC-LCNN performs better for the TTS attacks (A07–A16), whereas our method outperforms in challenging VC attacks (A17–A19). The performance is significantly improved for the unknown A17 (the most difficult attack for the baseline and top-performing challenge participants) and known A19 attacks, due to the strategic modifications introduced to the baseline RawNet2.

The superior performance on A17, A18, and A19, along with the competitive performance on TTS attacks (A07–A16), highlights the robustness of the proposed DeepRawNet for challenging spoofing attacks and consistent effectiveness. The evaluation conducted on both datasets shows that our approach is more proficient in locating the bonafide and spoofed audio and can effectively address key challenges encountered by the baseline models. The better feature engineering capability of our approach enables it to capture complicated patterns present in spoofed audio data, resulting in reduced EER and min-tDCF compared to the base models.

Comparison with existing methods

In this part, we have performed a thorough analysis of our approach with the latest approaches (*Wang & Yamagishi, 2023*, *2021*; *Yang et al., 2021*; *Lei et al., 2020*; *Li et al., 2021*; *Luo et al., 2021*; *Hua, Teoh & Zhang, 2021*; *Wang & Yamagishi, 2021*; *Kulkarni et al., 2024*; *Al-Tairi et al., 2025*) over the ASVspoof2019-LA, ASVspoof2021-LA, and ASVspoof2021-DF datasets, and results are provided in Table 14. *Wang & Yamagishi (2023)* investigated Neural Vocoders, *Wang & Yamagishi (2021)* employed the Wav2vec-LLGF approach, *Yang et al. (2021)* employed ResNet-18, light CNN (LCNN)

Table 13 Algorithm-wise comparison of the DeepRawNet approach with the base models over the ASVspoof2021-LA dataset.

Attacks	DeepRawNet (ours)		RawNet-2 (Baseline)		LFCC-LCNN (Baseline)		LFCC-GMM (Baseline)		CQCC-GMM (Baseline)	
	EER	min-tDCF	EER	min-tDCF	EER	min-tDCF	EER	min-tDCF	EER	min-tDCF
A07	1.87	0.1932	5.57	0.2838	0.80	0.1685	26.37	0.6822	9.17	0.3552
A08	7.34	0.3271	8.87	0.3817	5.42	0.2728	6.39	0.2849	2.94	0.2227
A09	2.44	0.4649	5.29	0.5678	0.41	0.3795	2.98	0.4190	1.43	0.4006
A10	1.92	0.1950	5.34	0.2789	0.80	0.1696	30.35	0.7544	12.06	0.4195
A11	2.53	0.2094	5.67	0.2843	0.78	0.1666	18.19	0.4942	8.81	0.3473
A12	2.64	0.2213	5.98	0.3036	0.57	0.1695	20.30	0.5473	11.30	0.4010
A13	1.31	0.1963	3.77	0.2628	0.66	0.1802	9.60	0.3712	10.19	0.3884
A14	3.11	0.2169	6.19	0.2890	0.54	0.1570	17.92	0.5070	7.50	0.3058
A15	2.55	0.2031	5.92	0.2807	0.41	0.1533	17.11	0.4900	6.38	0.2903
A16	2.89	0.2113	5.60	0.2768	1.42	0.1727	22.51	0.6065	16.92	0.5077
A17	15.92	0.9965	19.35	0.9999	23.62	0.9940	18.42	0.8679	29.02	0.9635
A18	25.97	0.9997	27.34	1.0000	29.27	0.9351	16.34	0.8661	24.24	0.9107
A19	4.96	0.4524	7.94	0.5448	14.73	0.5693	30.75	0.9314	43.20	0.9996
Average	5.80	0.3759	8.68	0.4426	6.11	0.3452	18.25	0.6017	14.09	0.5009

approach was suggested in *Yang et al. (2021)*, *Lei et al. (2020)* suggested the Siamese CNN framework, *Li et al. (2021)* proposed an SE-Res2Net50 DL approach, *Luo et al. (2021)* proposed a capsule network, and *Hua, Teoh & Zhang (2021)* introduced a Time-domain Synthetic Speech Detection Net (TSSDNet) approach having ResNet- or Inception-style structures for audio spoofing detection. The approaches (*Wang & Yamagishi, 2021*; *Kulkarni et al., 2024*) implemented HuBERt with LLGF and ECAPA models, respectively, while (*Al-Tairi et al., 2025*) suggested a RawNet2-based spoofing detection method. These approaches (*Wang & Yamagishi, 2023*; *2021*; *Yang et al., 2021*; *Lei et al., 2020*; *Li et al., 2021*; *Luo et al., 2021*; *Hua, Teoh & Zhang, 2021*; *Wang & Yamagishi, 2021*; *Kulkarni et al., 2024*; *Al-Tairi et al., 2025*) reported results over the ASVspoof2019-LA, ASVspoof2021-LA, and ASVspoof2021-DF datasets. The provided analysis in Table 14 shows that for the ASVspoof2019-LA dataset, the approaches (*Lei et al., 2020*; *Li et al., 2021*; *Luo et al., 2021*; *Hua, Teoh & Zhang, 2021*) indicated better performance in terms of min-tDCF and EER. Further, for the ASVspoof2021-LA dataset, we attained better results in terms of min-tDCF with a score of 0.3666, while for the EER measure, the work also obtained comparable results with a score of 7.77%. Further, for the ASVspoof2021-DF data sample, the works (*Lei et al., 2020*; *Li et al., 2021*; *Luo et al., 2021*; *Wang & Yamagishi, 2021*; *Hua, Teoh & Zhang, 2021*) attained better scores than our approach, with an EER of 20.72%. The results in Table 14 indicate that our proposed DeepRawNet exhibits strong and balanced performance across multiple benchmarks, effectively addressing the challenges of audio spoofing detection. In the ASVspoof2021-LA evaluation, our method achieves the lowest min-tDCF (0.3666) among all compared methods, demonstrating superior robustness to real-world transmission effects and codec variations. Additionally, it attains a lower EER of 7.77% compared to several existing approaches, highlighting its improved

Table 14 Performance comparison of the DeepRawNet with the latest approaches.

Methods	ASVspoof2019-LA		ASVspoof2021-LA		ASVspoof2021-DF
	min-tDCF	EER (%)	min-tDCF	EER (%)	EER (%)
Neural vocoders (*Wang & Yamagishi, 2023*)	—	2.98	—	7.53	23.62
Wav2vec-LLGF (*Wang & Yamagishi, 2021*)	—	3.45	—	10.97	20.75
ResNet (*Yang et al., 2021*)	0.197	6.49	0.6932	10.23	23.58
LCNN (*Yang et al., 2021*)	0.176	5.99	0.6811	9.68	21.36
Siamese CNN (*Lei et al., 2020*)	0.093	3.79	0.6525	8.59	20.01
SE-Res2Net50 (*Li et al., 2021*)	0.0743	2.50	0.5415	8.01	18.48
Capsule network (*Luo et al., 2021*)	0.0538	1.97	0.5028	7.35	17.26
Res-TSSDNet (*Hua, Teoh & Zhang, 2021*)	0.0481	1.64	0.4006	5.88	16.21
HuBERT–LLGF (*Wang & Yamagishi, 2021*)	–	3.55	–	9.55	13.07
HuBERT–Ecapa (*Kulkarni et al., 2024*)	–	1.05	–	12.55	13.79
DeepLASD (*Al-Tairi et al., 2025*)	0.1216	5.2753	0.4250	12.76	—
DeepRawNet (Proposed)	0.1275	2.80	0.3666	7.77	20.72

ability to differentiate between bonafide and spoofed audio. While methods in *Lei et al. (2020)*, *Li et al. (2021)*, *Luo et al. (2021)*, *Hua, Teoh & Zhang (2021)* achieve slightly better performance in certain cases, these models often come at the cost of increased complexity and a lack of generalization ability. In contrast, DeepRawNet maintains a favorable trade-off between detection accuracy and computational efficiency, making it more suitable for real-time and resource-constrained environments. Further, our proposed model being trained only on the ASVspoof2019-LA dataset, shows improved detection scores not only on the ASVspoof2019-LA dataset, but also on the ASVspoof2021-LA and DF data samples (Table 14), indicating the effective generalization power of the model. These results validate that DeepRawNet introduces an optimized and effective architecture for generalized and robust spoof detection, ensuring high reliability across diverse datasets and real-world deployment scenarios.

Evaluation under acoustic distortions

To evaluate the robustness of the proposed DeepRawNet under noisy and acoustic distortions, we conducted an experiment using the perturbed versions of the evaluation set of the ASVspoof2019-LA dataset. Specifically, perturbations, including Gaussian noise, pitch shift, time stretch, and temporal shift, were applied individually to evaluation utterances, resulting in the perturbed versions of the dataset. The proposed model trained on the train set of ASVspoof2019-LA was then evaluated using the perturbed versions of the unseen evaluation set of ASVspoof2019-LA, and the results are reported in Table 15. The increased values of min-tDCF and EER demonstrate that the perturbations negatively impact the performance. Particularly, time stretch and pitch shift caused significant performance degradation, which indicates the reliance of the model on spectral-temporal consistency while detecting the spoofed audio. The temporal shift caused comparatively less reduction in performance; however, the Gaussian noise caused the least performance

Table 15 Performance under acoustic distortions using ASVspoof-2019 LA dataset.

Audio perturbations	min-tDCF	EER (%)
Gaussian noise	0.1790	5.20
Pitch shift	0.8047	24.31
Time stretch	0.7796	23.63
Temporal shift	0.3160	11.82

Table 16 Performance comparison with step-by-step modifications added in RawNet2.

Model modifications			ASVspoof2019-LA		ASVspoof2021-LA	
Transpose conv	PReLU	Log softmax	min-tDCF	EER (%)	min-tDCF	EER (%)
✓			0.2228	7.04	0.4901	12.16
	✓		0.1672	4.92	0.3884	9.86
		✓	0.1866	5.56	0.4209	10.45
✓	✓	✓	0.1275	2.80	0.3666	7.77

drop, signifying that the proposed model can detect manipulation artifacts reliably under noisy environments. In general, the attained results highlight the proposed model's robustness for the background noise. However, it reveals that the speech transformations affecting the temporal and spectral features of the audio signal reduce the performance for spoofing detection.

Ablation study

In the subsequent experiment, we conducted an ablation study to analyze the impact of different architectural components on the performance and generalizability of our model by evaluating it on the ASVspoof2019-LA and ASVspoof2021-LA datasets. For this, we conducted 2 types of experiments. In the first experiment, we evaluated the impact of individual architectural modifications, and attained results are reported in Table 16. In the first phase, we only integrated Transpose Convolutions in the base RawNet2 model. In the next phase, we have checked the impact of adding the PReLU activation function. Lastly, we have analyzed the impact of using Log Softmax in the final classification layer. It can be observed from the results in Table 16 that while each modification individually contributed to performance improvements, the combination of all three yielded the best results. This demonstrates that the joint impact of these modifications creates a synergistic effect by addressing different aspects of model optimization effectively.

In the second experiment, we compared our DeepRawNet approach against various configurations of RawNet2, including RawNet2 with ELU, ReLU, and Hardswish activation functions, RawNet2 with RNN, and an LSTM layer (Table 17). By systematically varying these components and evaluating their performance on the ASVspoof2019-LA and ASVspoof2021-LA datasets, we aimed to identify the most effective architecture for deepfake audio detection. This ablation study provides valuable insights into the contribution of each architectural element to the overall performance of our model,

Table 17 Performance comparison with various configurations of RawNet2.

Models	ASVspoof2019-LA dataset		ASVspoof2021-LA dataset	
	EER (%)	min-tDCF	EER (%)	min-tDCF
RawNet2-ELU	5.57	0.1832	10.95	0.4312
RawNet2-ReLU	5.82	0.1841	11.31	0.4481
RawNet2-Hardswish	6.61	0.2072	11.69	0.4855
RawNet2-RNN	20.73	0.4049	28.95	0.7141
RawNet2-LSTM	7.95	0.2391	17.32	0.5385
DeepRawNet (Proposed)	2.80	0.1275	7.77	0.3666

helping to refine and optimize its design for enhanced performance. In the ablation study, we observed that our proposed DeepRawNet consistently outperformed the other configurations across all evaluation metrics on the ASVspoof2019-LA and ASVspoof2021-LA datasets. While conducting the ablation study, we also observed certain limitations with the other configurations of RawNet2. For instance, the variant utilizing the ELU activation exhibited suboptimal performance, possibly due to its inability to effectively handle dead neurons and capture complex patterns in raw audio. Similarly, the configurations incorporating ReLU and Hardswish activation functions also faced challenges in effectively extracting discriminative features, leading to higher EER and min t-DCF values.

Additionally, the variants incorporating RNN and LSTM layers showed limited improvement in performance, indicating that these recurrent architectures may not be well-suited for capturing long-range dependencies in raw audio. In comparison, the suggested DeepRawNet approach performs better than all other configurations of the RawNet. Specifically, the variant incorporating the PReLU activation, and the transpose convolution layer in the residual block exhibited superior performance. The ablation study demonstrates that the contribution of each component is consistent across both ASV datasets. This underscores the effectiveness of our proposed enhancements in improving our model's ability to capture discriminative patterns in raw audio for reliable deepfake detection compared to alternative configurations.

DeepRawnet-verifier tool

We propose a robust and effective voice spoofing detection tool named DeepRawNet-Verifier to combat the prevailing threats related to audio deepfakes. As the audio deepfake techniques continue to evolve rapidly, there is a need to develop reliable systems capable of discriminating between spoofed and bonafide audio. The DeepRawNet-Verifier tool is designed to accurately detect the fake audio generated using VC and TTS techniques, thus helping to preserve the integrity of audio content shared on media platforms. The proposed tool has a simple and user-friendly interface, making navigation and utilization easy for researchers and practitioners. The interface of the DeepRawNet-Verifier tool is shown in Fig. 4.

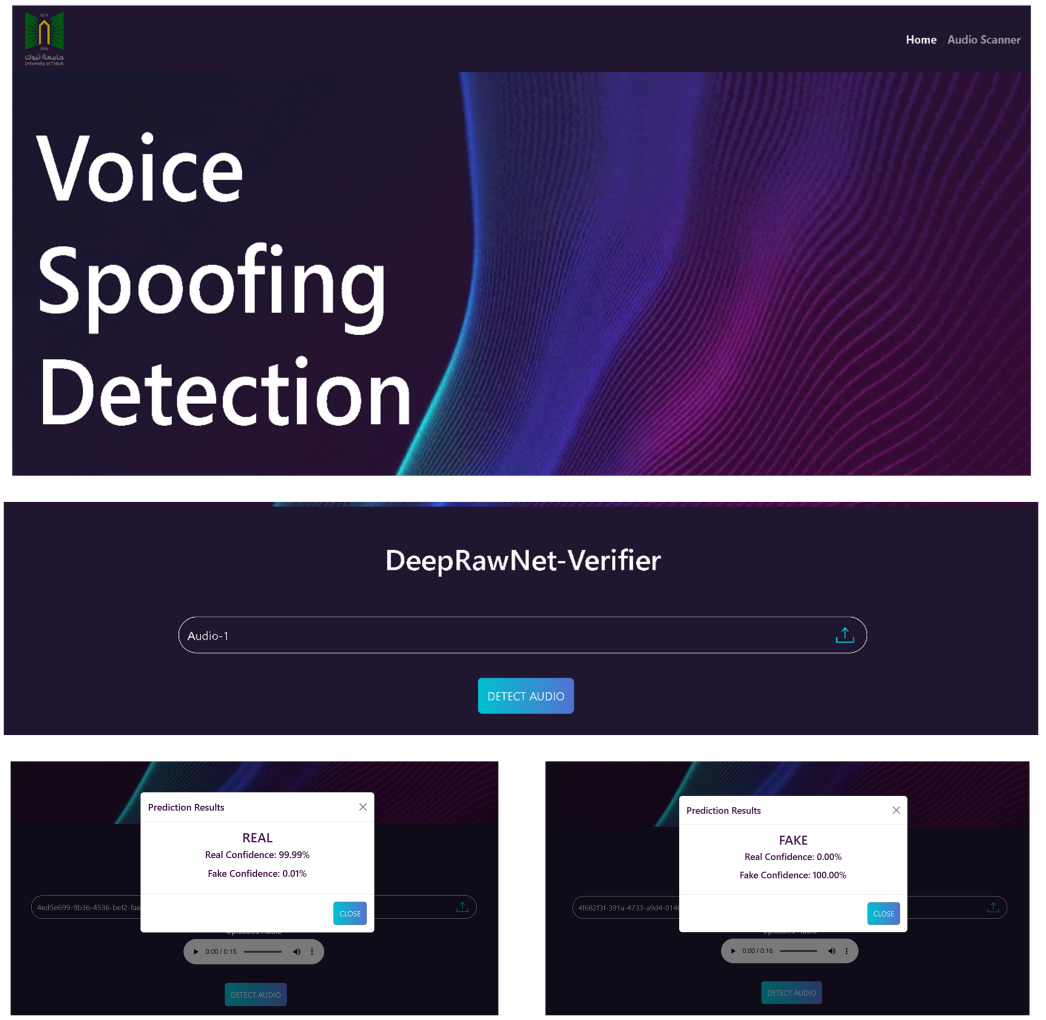

Figure 4 Proposed DeepRawNet-verifier tool interface. Full-size DOI: 10.7717/peerj-cs.3670/fig-4

DeepRawNet- Verifier provides an interactive pipeline for detecting spoofed audio, integrating the frontend interface with the backend engine. The complete workflow involves audio selection, detection trigger, audio analysis, and result display.

Audio Selection: The user can initiate the process by selecting an audio file through the DeepRawNet-Verifier tool interface.

Detection Trigger: After that, the user can trigger the detection process by clicking the "Detect Audio" button, which initiates the API call and processes the audio file at the backend.

Audio Analyzing: The tool's backend is powered by the proposed DeepRawNet framework, which analyzes and detects the given audio sample. The predicted output is then returned to the front end of the tool.

Result Display: Finally, the prediction results returned by the backend are displayed on the frontend, including the label and prediction scores indicating the proposed model's certainty on its prediction.

It is important to note that currently, our tool implements the proposed DeepRawNet framework; however, in the future, we intend to scale the tool *via* integrating more audio spoofing detection models to further enhance the reliability and trustworthiness of the DeepRawNet-Verifier tool.

DISCUSSION

The presented DeepRawNet method reliably recognizes the audio spoofing attacks, which we have proved through accomplishing numerous experiments on standard datasets named the ASVspoof-2019-LA, 2021-LA, and 2021-DF datasets, respectively. Initial experiments involved training the model on the ASV2019-LA dataset and evaluating its performance on the unseen evaluation set. Results indicated a notable performance improvement with min-tDCF and EER of 0.1275% and 2.80% when the model was trained using the train and development set as training and validation data, respectively. Evaluation on the ASV2021-LA and ASV2021-DF datasets demonstrated consistent performance across different datasets, highlighting the model's robustness and generalization ability. Specifically, when the model is evaluated on the ASV2021-LA-Eval set, we got the min-tDCF and EER of 0.3666 and 7.77%, while for the ASV2021-DF-Eval set, our DeepRawNet approach attained the EER of 20.72%. Comparative analysis with baseline models such as LFCC-LCNN, LFCC-GMM, CQCC-GMM, and RawNet2 revealed significant improvements in min-tDCF and EER. Experimental results over both sets indicate that our approach is effective in handling the compression, codec, and transmission channel variability, even though not trained under these conditions. Further, we designed the ablation study section, which provides insights into the impact of the results obtained by our DeepRawNet approach against different architectural components and various configurations of RawNet2. The attained results reveal that our proposed DeepRawNet consistently outperformed other configurations across all evaluation metrics on the ASVspoof2019-LA dataset by achieving the min-tDCF and EER of 0.1275 and 2.80%, respectively.

Furthermore, the algorithm-wise evaluation allowed assessment of the proposed approach against specific attacks from both datasets, with promising results despite challenges posed by attacks, such as A17 and A18. The improved performance for A17 attacks can be attributed to the specific enhancements introduced in our DeepRawNet architecture. By increasing the negative slope in the Fixed Sinc filters and upgrading the activation functions to PReLU in residual blocks, our model demonstrates superior adaptability and feature extraction capabilities. These modifications allow our model to effectively capture the noise characteristic of A17 attacks. Furthermore, our DeepRawNet architecture enhances the representation and discrimination of phase-related artifacts present in the A17 attack. Similarly, for the A18 attack, which exhibits different acoustic properties compared to other attacks, our DeepRawNet model is equipped with enhanced adaptability to diverse input distributions. By dynamically responding to the unique

characteristics of the A18 attack, the model can extract discriminative features more accurately, leading to improved detection performance.

Further, we designed an experiment where we evaluated the performance of the DeepRawNet over the VC samples from the ASVspoof2019-LA and ASVspoof2021-LA datasets, which closely resemble genuine speech. Our method, when tested on a diverse range of VC samples, showed improved performance in accurately detecting these challenging attacks, as evidenced by significant reductions in EER and min-tDCF compared to the baseline methods over both datasets. This underscores the effectiveness of our approach in addressing the unique challenges posed by VC samples and advancing deepfake audio detection capabilities. Finally, in the last experiment, we checked the results of our approach on the TTS samples from both datasets. For the TTS samples from the ASVspoof2021-LA dataset, we performed well in comparison to the baseline approaches, which signifies the effectiveness of the DeepRawNet to handle the codec and transmission variations. Whereas, for the TTS samples from ASVspoof2019-LA, we attained improved results in comparison with baseline approaches. A little performance degradation has been observed over the RawNet2, which emphasizes the need for further fine-tuning and possibly incorporating additional techniques, such as ensemble learning, to better adapt our model for TTS detection. In general, the proposed approach represents a significant advancement in deepfake audio detection, offering improved accuracy, robustness, and generalization across diverse datasets and attack types, and can effectively handle the codec, transmission variations, and compressed samples. Future research directions include exploring novel architectures, incorporating additional features, and investigating advanced training techniques to further enhance the model's performance and address emerging threats in audio spoofing.

CONCLUSIONS

This work has presented a DeepRawNet method that provides a significant advancement in deepfake audio detection. Through strategic enhancements, our DeepRawNet exhibits improved adaptability and robustness in capturing complex patterns in raw audio, thus making it a more effective solution for spoofing detection tasks. We evaluated the performance of the DeepRawNet on ASVspoof2019-LA and ASVspoof2021-LA/DF datasets and attained improved results over the baseline approaches, including the original RawNet2 model, and better recognized the more complex VC samples. Collectively, the proposed approach exhibits a significant advancement in recognizing audio deepfakes, offering improved accuracy, robustness, and generalization across diverse datasets and attack types, and can effectively handle the codec, transmission variations, and compressed samples.

Limitations and future work

Although the proposed DeepRawNet framework demonstrates promising performance, several limitations should be acknowledged. DeepRawNet yields marginally lower results on TTS samples compared to RawNet2, indicating a need for further fine-tuning and the potential use of ensemble methods. Additionally, its evaluation under real-world acoustic

conditions and robustness against adversarial attacks need to be further explored. Furthermore, the evaluation is limited to ASV datasets that provide a solid benchmark; however unable to comprehensively represent the diverse real-world conditions, multilingual, and cross-lingual spoofing scenarios. Looking forward, future research directions involve refining RawNet2's architecture to handle the challenges in deepfake audio detection, and more comprehensive benchmarking by evaluating the framework on in-the-wild and multilingual spoofing datasets. Furthermore, future work will also incorporate novel feature extraction techniques from raw audio waveforms, noise-robust feature processing, environmental data augmentation, and domain adaptation strategies to better assess and enhance the model's generalization in real-world scenarios.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the Deanship of Research and Graduate Studies at the University of Tabuk through Research No. S-1444-0108. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Deanship of Research and Graduate Studies at the University of Tabuk: S-1444-0108.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Lubna A. Alharbi conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, authored or reviewed drafts of the article, funding Acquisition, and approved the final draft.
- Jasim Alnahas conceived and designed the experiments, performed the experiments, performed the computation work, authored or reviewed drafts of the article, and approved the final draft.
- Ali Javed conceived and designed the experiments, performed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- A'aeshah Alhakamy conceived and designed the experiments, prepared figures and/or tables, and approved the final draft.
- Marriam Nawaz conceived and designed the experiments, analyzed the data, prepared figures and/or tables, and approved the final draft.
- Hafiz Malik conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Hafsa Ilyas performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, and approved the final draft.
- Faris Alfaifi performed the experiments, performed the computation work, prepared figures and/or tables, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The code is available in the Supplemental File.

The ASVspoof 2019 dataset is available at: https://datashare.ed.ac.uk/handle/10283/3336.

The ASVspoof 2021 dataset is available at: https://www.asvspoof.org/index2021.html.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.3670#supplemental-information.

REFERENCES

Abdelhamid AA, El-Kenawy ESM, Alotaibi B, Amer GM, Abdelkader MY, Ibrahim A, Eid MM. 2022. Robust speech emotion recognition using CNN+ LSTM based on stochastic fractal search optimization algorithm. *IEEE Access* **10**:49265–49284 DOI 10.1109/ACCESS.2022.3172954.

Al-Tairi H, Javed A, Khan T, Saudagar AKJ. 2025. DeepLASD countermeasure for logical access audio spoofing. *Scientific Reports* **15(1)**:20839 DOI 10.1038/s41598-025-04808-5.

Almutairi Z, Elgibreen H. 2022. A review of modern audio deepfake detection methods: challenges and future directions. *Algorithms* **15(5)**:155 DOI 10.3390/a15050155.

Alsalhi A, Almehmadi A. 2024. Using vocal-based emotions as a human error prevention system with convolutional neural networks. *Applied Sciences* **14(12)**:5128 DOI 10.3390/app14125128.

Borodin K, Kudryavtsev V, Korzh D, Efimenko A, Mkrtchian G, Gorodnichev M, Rogov OY. 2024. AASIST3: KAN-enhanced AASIST speech deepfake detection using SSL features and additional regularization for the ASVspoof 2024 challenge. ArXiv DOI 10.48550/arXiv.2408.17352.

Boyd J, Fahim M, Olukoya O. 2023. Voice spoofing detection for multiclass attack classification using deep learning. *Machine Learning with Applications* **14**:100503 DOI 10.1016/j.mlwa.2023.100503.

Chaabane SB, Hijji M, Harrabi R, Seddik H. 2022. Face recognition based on statistical features and SVM classifier. *Multimedia Tools and Applications* **81(6)**:8767–8784 DOI 10.1007/s11042-021-11816-w.

Chakravarty N, Dua M. 2023. Spoof detection using sequentially integrated image and audio features. *International Journal of Computing Digital Systems* **13(1)**:1.

Chakravarty N, Dua M. 2024. A lightweight feature extraction technique for deepfake audio detection. *Multimedia Tools Applications* **83**:67443–67467 DOI 10.1007/s11042-024-18217-9.

Cohen A, Rimon I, Aflalo E, Permuter HH. 2022. A study on data augmentation in voice anti-spoofing. *Speech Communication* **141**:56–67 DOI 10.1016/j.specom.2022.04.005.

Cuccovillo L, Papastergiopoulos C, Vafeiadis A, Yaroshchuk A, Aichroth P, Votis K, Tzovaras D. 2022. Open challenges in synthetic speech detection. In: *2022 IEEE International Workshop on Information Forensics and Security (WIFS)*. Piscataway: IEEE.

Das RK. 2021. Known-unknown data augmentation strategies for detection of logical access, physical access and speech deepfake attacks: ASVspoof 2021. In: *Proceeding Edition of the Automatic Speaker Verification Spoofing Countermeasures Challenge*, 29–36.

Dawood H, Saleem S, Hassan F, Javed A. 2022. A robust voice spoofing detection system using novel CLS-LBP features and LSTM. *Journal of King Saud University-Computer Information Sciences* **34(9)**:7300–7312 DOI 10.1016/j.jksuci.2022.02.024.

Dişken G. 2024. Complementary regional energy features for spoofed speech detection. *Computer Speech Language* **85**:101602 DOI 10.1016/j.csl.2023.101602.

Doan TP, Hong K, Jung S. 2023. GAN discriminator-based audio deepfake detection. In: *Proceedings of the 2nd Workshop on Security Implications of Deepfakes and Cheapfakes*, 29–32.

Dua M, Chakravarty N, Reddy SGP, Bansal A, Pawar S, Dua S. 2025. MelCochleaGram-DeepCNN: sequentially fused spectrogram and the DeepCNN classifiers-based audio spoof detection system. *IETE Journal of Research* **71(1)**:65–70 DOI 10.1080/03772063.2024.2412799.

Ge W, Wang X, Yamagishi J, Todisco M, Evans N. 2024. Spoofing attack augmentation: can differently-trained attack models improve generalisation? In: *ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Piscataway: IEEE, 12531–12535.

Habbash M, Mnasri S, Alghamdi M, Alrashidi M, Tarawneh AS, Gumair A, Hassanat AB. 2024. Recognition of Arabic accents from English spoken speech using deep learning approach. *IEEE Access* **12**:37219–37230 DOI 10.1109/ACCESS.2024.3374768.

Hamza A, Javed AR, Iqbal F, Kryvinska N, Almadhor AS, Jalil Z, Borghol R. 2022. Deepfake audio detection via MFCC features using machine learning. *IEEE Access* **10**:134018–134028 DOI 10.1109/ACCESS.2022.3231480.

Hijji M, Alam G. 2021. A multivocal literature review on growing social engineering based cyber-attacks/threats during the COVID-19 pandemic: challenges and prospective solutions. *IEEE Access* **9**:7152–7169 DOI 10.1109/ACCESS.2020.3048839.

Hua G, Teoh ABJ, Zhang H. 2021. Towards end-to-end synthetic speech detection. *IEEE Signal Processing Letters* **28**:1265–1269 DOI 10.1109/LSP.2021.3089437.

Huang W, Gu Y, Wang Z, Zhu H, Qian Y. 2025. Generalizable audio deepfake detection via latent space refinement and augmentation. In: *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Piscataway: IEEE.

Huang L, Pun C-M. 2024. Self-attention and hybrid features for replay and deep-fake audio detection. ArXiv DOI 10.48550/arXiv.2401.05614.

Iqbal F, Abbasi A, Javed AR, Jalil Z, Al-Karaki J. 2022. Deepfake audio detection via feature engineering and machine learning. In: *Woodstock'22: Symposium on the Irreproducible Science*. Piscataway: IEEE, 70–75.

Javed A, Malik KM, Malik H, Irtaza A. 2022. Voice spoofing detector: a unified anti-spoofing framework. *Expert Systems with Applications* **198**:116770 DOI 10.1016/j.eswa.2022.116770.

Jung JW, Heo HS, H.Tak HJS, Chung JS, Lee BJ, Yu HJ, Evans N. 2022. AASIST: audio anti-spoofing using integrated spectro-temporal graph attention networks. In: *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Piscataway: IEEE, 6367–6371.

Jung JW, Heo HS, Kim JH, Shim HJ, Yu HJ. 2019. RawNet: advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification. ArXiv DOI 10.48550/arXiv.1904.08104.

Kassis A, Hengartner U. 2021. Practical attacks on voice spoofing countermeasures. ArXiv DOI 10.48550/arXiv.2107.14642.

Khan A, Malik KM, Ryan J, Saravanan M. 2023a. Voice spoofing attacks and countermeasures: a systematic review, analysis, and experimental evaluation. *Research Square* DOI 10.21203/rs.3.rs-2557691/v1.

Khan A, Malik KM, Ryan J, Saravanan M. 2023b. Battling voice spoofing: a review, comparative analysis, and generalizability evaluation of state-of-the-art voice spoofing counter measures. *Artificial Intelligence Review* **56(Suppl 1)**:513–566 DOI 10.1007/s10462-023-10539-8.

Kulkarni A, Tran HM, Kulkarni A, Dowerah S, Lolive D, Doss MM. 2024. Exploring generalization to unseen audio data for spoofing: insights from SSL models. In: *ASVSpoof Workshop*.

Lei Z, Yang Y, Liu C, Ye J. 2020. Siamese convolutional neural network using gaussian probability feature for spoofing speech detection. In: *Interspeech*, 1116–1120.

Li X, Li N, C.Weng XL, Su D, Yu D, Meng H. 2021. Replay and synthetic speech detection with Res2Net architecture. In: *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Piscataway: IEEE, 6354–6358.

Li L, Lu T, Ma X, Yuan M, Wan D. 2023. Voice deepfake detection using the self-supervised pre-training model HuBERT. *Applied Sciences* **13(14)**:8488 DOI 10.3390/app13148488.

Liu X, Sahidullah M, Lee KA, Kinnunen T. 2024. Generalizing speaker verification for spoof awareness in the embedding space. *IEEE/ACM Transactions on Audio, Speech, Language Processing* **32**:1261–1273 DOI 10.1109/TASLP.2024.3358056.

Liu R, Zhang J, Gao G. 2024. Multi-space channel representation learning for mono-to-binaural conversion based audio deepfake detection. *Information Fusion* **105**:102257 DOI 10.1016/j.inffus.2024.102257.

Liz-López H, Keita M, Taleb-Ahmed A, Hadid A, Huertas-Tato J, Camacho D. 2024. Generation and detection of manipulated multimodal audiovisual content: advances, trends and open challenges. *Information Fusion* **103**:102103 DOI 10.1016/j.inffus.2023.102103.

Luo A, Li E, Liu Y, Kang X, Wang ZJ. 2021. A capsule network based approach for detection of audio spoofing attacks. In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Piscataway: IEEE, 6359–6363.

Ma Q, Zhong J, Yang Y, Liu W, Gao Y, Ng W. 2023. A lightweight and efficient model for audio anti-spoofing. In: *Proceedings of the 5th ACM International Conference on Multimedia in Asia*, 1–7.

Masood M, Nawaz M, Malik KM, Javed A, Irtaza A, Malik H. 2023. Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. *Applied Intelligence* **53(4)**:3974–4026 DOI 10.1007/s10489-022-03766-z.

Mastromichalakis S. 2020. ALReLU: a different approach on leaky ReLU activation function to improve neural networks performance. ArXiv DOI 10.48550/arXiv.2012.07564.

Meriem F, Messaoud B, Bahia YZ. 2023. Texture analysis of edge mapped audio spectrogram for spoofing attack detection. *Multimedia Tools Applications* **83(14)**:1–23 DOI 10.1007/s11042-023-15329-6.

Neelima M, Prabha IS. 2023. Optimized deep network based spoof detection in automatic speaker verification system. *Multimedia Tools Applications* **83**:13073–13091 DOI 10.1007/s11042-023-16127-w.

Nguyen-Vu L, Doan T-P, Bui M, Hong K, Jung S. 2023. On the defense of spoofing countermeasures against adversarial attacks. *IEEE Access* **11**:94563–94574 DOI 10.1109/ACCESS.2023.3310809.

QingJie W, WenBin W. 2017. Research on image retrieval using deep convolutional neural network combining L1 regularization and PRelu activation function. In: *IOP Conference Series: Earth and Environmental Science*. Vol. 69, IOP Publishing, 012156.

RawNet2 Code. 2021. RawNet2 ASVspoof 2021 baseline. *Available at https://github.com/asvspoof-challenge/2021/tree/main/LA/Baseline-RawNet2* (accessed 22 October 2025).

Rishith Sadashiv TN, Bedge A, Bore SS, Mishra J, Bhattacharjee M, Prasanna SM. 2025. Fusion of modulation spectrogram and SSL with multi-head attention for fake speech detection. ArXiv DOI 10.48550/arXiv.2508.01034.

Sinha S, Dey S, Saha G. 2024. Improving self-supervised learning model for audio spoofing detection with layer-conditioned embedding fusion. *Computer Speech Language* **86**:101599 DOI 10.1016/j.csl.2023.101599.

Stergiou A, Poppe R, Kalliatakis G. 2021. Refining activation downsampling with SoftPool. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10357–10366.

Tak H, Patino J, Todisco M, Nautsch A, Evans N, Larcher A. 2021. End-to-end anti-spoofing with RawNet2. In: *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Piscataway: IEEE, 6369–6373.

Wang X, Yamagishi J. 2021. Investigating self-supervised front ends for speech spoofing countermeasures. ArXiv DOI 10.48550/arXiv.2111.07725.

Wang X, Yamagishi J. 2023. Spoofed training data for speech spoofing countermeasure can be efficiently created using neural vocoders. In: *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Piscataway: IEEE.

Wei J, Yao L, Meng Q. 2023. Self-adaptive logit balancing for deep neural network robustness: defence and detection of adversarial attacks. *Neurocomputing* **531**:180–194 DOI 10.1016/j.neucom.2023.02.013.

Xue J, Fan C, Lv Z, Tao J, Yi J, Zheng C, Wen Z, Yuan M, Shao S. 2022. Audio deepfake detection based on a combination of F0 information and real plus imaginary spectrogram features. In: *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, 19–26.

Xue J, Zhou H. 2023. Physiological-physical feature fusion for automatic voice spoofing detection. *Frontiers of Computer Science* **17(2)**:172318 DOI 10.1007/s11704-022-2121-6.

Yang J, Chen F, Cheng Y, Lin P. 2024. Integration of audio-visual information for multi-speaker multimedia speaker recognition. *Digital Signal Processing* **145**:104315 DOI 10.1016/j.dsp.2023.104315.

Yang T, Wang H, Das RK, Qian Y. 2021. Modified magnitude-phase spectrum information for spoofing detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **29**:1065–1078 DOI 10.1109/TASLP.2021.3060810.

Yu Z, Chang Y, Zhang N, Xiao C. 2023. SMACK: semantically meaningful adversarial audio attack. In: *32nd USENIX Security Symposium (USENIX Security 23)*, 3799–3816.

Zhang Y, Lu J, Shang Z, Wang W, Zhang P. 2024. Improving short utterance anti-spoofing with AASIST2. In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Piscataway: IEEE, 11636–11640.

Zhang J, Tu G, Liu S, Cai Z. 2023. Audio anti-spoofing based on audio feature fusion. *Algorithms* **16(7)**:317 DOI 10.3390/a16070317.

Zhou J, Hai T, Jawawi DN, Wang D, Ibeke E, Biamba C. 2022. Voice spoofing countermeasure for voice replay attacks using deep learning. *Journal of Cloud Computing* **11(1)**:51 DOI 10.1186/s13677-022-00306-5.