Adversarial Attacks on Audio Deepfake Detection: A Benchmark and Comparative Study

Kutub Uddin
kutub@umich.edu
Muhammad Umar Farooq
mufarooq@umich.edu
Awais Khan
mawais@umich.edu
Khalid Mahmood Malik
drmalik@umich.edu

College of Innovation & Technology,
University of Michigan-Flint,
MI, 48502, USA

Abstract

The widespread use of generative AI has shown remarkable success in producing highly realistic deepfakes, posing a serious threat to various voice biometric applications, including speaker verification, voice biometrics, audio conferencing, and criminal investigations. To counteract this, several state-of-the-art (SoTA) audio deepfake detection (ADD) methods have been proposed to identify generative AI signatures to distinguish between real and deepfake audio. However, the effectiveness of these methods is severely undermined by anti-forensic (AF) attacks that conceal generative signatures. These AF attacks span a wide range of techniques, including statistical modifications (e.g., pitch shifting, filtering, noise addition, and quantization) and optimization-based attacks (e.g., FGSM, PGD, C & W, and DeepFool). In this paper, we investigate the SoTA ADD methods and provide a comparative analysis to highlight their effectiveness in exposing deepfake signatures, as well as their vulnerabilities under adversarial conditions. We conducted an extensive evaluation of ADD methods on five deepfake benchmark datasets using two categories: raw and spectrogram-based approaches. This comparative analysis enables a deeper understanding of the strengths and limitations of SoTA ADD methods against diverse AF attacks. It does not only highlight vulnerabilities of ADD methods, but also informs the design of more robust and generalized detectors for real-world voice biometrics. It will further guide future research in developing adaptive defense strategies that can effectively counter evolving AF techniques.

1 Introduction

The rapid advancement of generative AI has significantly increased the accessibility of deepfakes [31] to mimic human voices. Although deepfakes improve user convenience in voice-biometrics, including smart assistants, such as Amazon Alexa, Google Home, and Apple Siri, they also introduce serious risks [21]. In particular, attackers can use deepfakes to impersonate users and bypass voice biometrics [15]. Consequently, systems ranging from automatic speaker verification to ADDs are facing security challenges. Recent analyses from industry and academia reports reveal a sharp increase in deepfake-related incidents [25]. In

2024 alone, an estimated $12.5 billion was lost to fraud related to AI-driven attacks, with approximately $2.6 million of that loss specifically attributed to audio deepfakes [24].

To address these risks, the research community has proposed several ADDs [14]. These ADDs are based on raw audio or acoustic features (e.g., MFCC, LFCC, Spectrogram) with traditional classifiers [12, 13, 16], end-to-end deep learning [10, 28, 29], and more recent fine-tuned foundation models [14]. In particular, ADDs aim to capture subtle artifacts from raw or spectrogram signals during generation. However, as ADDs become more sophisticated, they have become increasingly vulnerable to anti-forensics (AFs). AFs span a diverse techniques, from statistical (e.g., noise injection [17]), to recent optimization (e.g., FGSM [7], PGD [19], C & W [2], and DeepFool [20]) that can significantly degrade the performance of ADDs.

Although robustness has been extensively explored in the image domain [1, 23, 32, 33, 34, 36], its application to ADD remains relatively unexplored. Some recent studies [5, 11, 27, 37, 42] have investigated the vulnerability of ADDs and demonstrated the performance degradation by AFs. For example, Wu et al. [42] presented contrastive learning-based ADD to improve robustness against AFs, such as volume control, fading, and noise injection. Although effective against these perturbations, the study [42] does not address complex AFs such as optimization [7, 35]. Similarly, studies such as [5, 11] are limited in both architectural scope and dataset diversity. Furthermore, a recent survey [27] reviews AFs on audiovisual deepfake, including defense methods such as fusion and decoy techniques, but lacks comparative evaluations of vulnerabilities against AF attacks.

Despite progress [5, 27], a critical gap remains in unified evaluation of ADDs under diverse AFs across various designs and datasets. Most ADDs [10, 15, 28, 29] focus on limited perturbations (e.g., noise) or test on selected datasets without cross-corpus evaluation. Thus, current evaluations miss AF scenarios where attackers exploit ADDs to evade them.

To our knowledge, this is the first large-scale comparative study of AFs on both raw and spectrogram-based ADDs across diverse datasets and AF types. Due to the lack of standardized evaluation, it remains unclear which methods are more robust. We address this gap through a unified, cross-architecture, and cross-dataset analysis to offer a comprehensive study across input formats, ADDs, and AFs. Our key contributions are:

- We present the first unified, large-scale, and apple-to-apple comparative evaluation of SoTA ADDs under AF attacks, focused on voice biometric applications.
- We benchmark twelve ADDs across five datasets under two AF categories: statistical (e.g., noise) and optimization-based (e.g., FGSM, PGD, C & W, DeepFool) attacks.
- We provide an in-depth analysis of current methods regarding key limitations and suggesting insights for designing more robust and generalizable ADD methods.

2 Benchmarks Selection

This section describes the datasets, ADDs, and AF techniques for the comparative study.

2.1 Datasets

Based on the recent benchmark study, we selected five widely used datasets to assess the performance and vulnerabilities of ADD systems. The selected datasets are as follows:

ASVSpoof2019 (*D*1) [31]: The ASVSpoof2019 dataset [31] includes logical and physical access tracks. We use a logical access subset that contains over 121,000 utterances generated using 17 TTS and VC systems.

ASVSpoof2021 (*D*2) [18]: The ASVSpoof2021 dataset [18] builds on previous editions [31] with a wider set of spoofing attacks on logical and physical access tracks and comprises more than 181,000 utterances.

ASVSpoof2024 (*D3*) [39]: The ASVSpoof2024 dataset [39] is the most recent in the ASVSpoof series, introducing multilingual, cross-device, and adversarial spoofing conditions. It comprises 182,000 training, 140,000 development, and 680,000 test samples.
CodecFake (*D4*) [43]: The CodecFake dataset [43] includes deepfakes generated by Audio Language Models (ALMs), which combine language modeling and codec compression to synthesize high-fidelity speech.
WaveFake (*D5*) [6]: The WaveFake dataset [6] is a large-scale dataset that includes over 117,000 clips synthesized using state-of-the-art vocoders and GANs, offering diverse generation techniques for ADD benchmarking.

2.2 Deepfake Detection Methods

We consider two principal categories of SoTA ADD models: (a) raw audio and (b) spectrogram-based methods. These methods were chosen based on their open-source availability, proven effectiveness, and representation of diverse architectural designs.

2.2.1 Raw Audio-based Methods

Raw ADDs operate directly on time-domain signals, learning temporal and low-level patterns to distinguish real and deepfakes. In our benchmark, we include six SoTA raw ADDs, including RawNet3 (*R1*) [10], MS-ResNet (*R2*) [38], SeNet (*R3*) [41], LCNN (*R4*) [11], Res-TSSDNet (*R5*) [9], and Inc-TSSDNet (*R6*) [9]. In particular, RawNet3 (*R1*) [10] integrates the Res2Net backbone and multi-layer feature aggregation, which shows strong performance even with limited labeled data. MS-ResNet (*R2*) [38] and SENet (*R3*) [41] are both ResNet-based variants; the MS-ResNet (*R2*) [38] incorporates multi-scale residual blocks to extract spatial patterns, while the SENet (*R3*) [41] enhances robustness via adversarial training and spatial smoothing. LCNN (*R4*) [11] introduces lightweight convolutional layers to

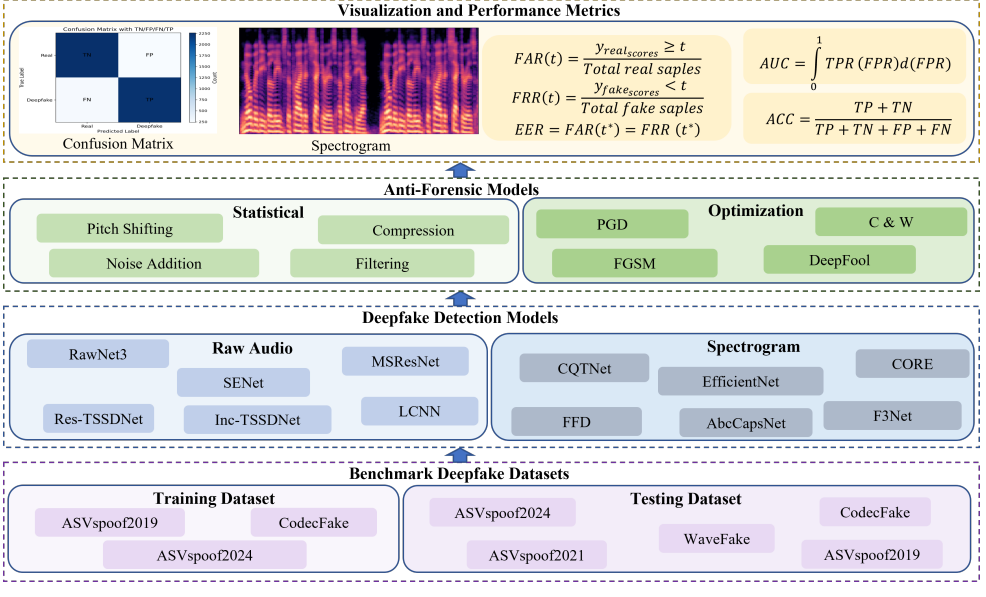


Figure 1: Architecture of the proposed comparative study. It starts with dataset selection, followed by two ADD groups, such as raw and spectrogram-based, to reveal synthetic traits. Two AF attack types, such as statistical and optimization-based, are applied to deceive the ADDs. Visualizations offer perceptibility and qualitative analysis across AF attack types.

learn discriminative features from raw waveforms. Res-TSSDNet (*R5*) [9] and Inc-TSSDNet (*R6*) [9] are lightweight, end-to-end models optimized for ADD. Res-TSSDNet (*R5*) [9] uses a ResNet-style architecture with stacked residual blocks, whereas Inc-TSSDNet (*R6*) [9] employs an Inception with parallel dilated convolutions to expand receptive fields. Both demonstrate strong generalization across datasets.

2.2.2 Spectrogram-based Methods

Spectrogram-based ADDs first transform audio into time-frequency (e.g., Mel or CQT spectrograms), then capture visual patterns to detect deepfake artifacts. Our benchmark includes six SoTA spectrogram-based ADDs, including ABCCapsNet (*S1*) [40], EfficientNet (*S2*) [30], FFD (*S3*) [4], CORE (*S4*) [22], F3Net (*S5*) [26], and CQTNet (*S6*) [45].
ABCCapsNet (*S1*) [40] combines attention-based bottleneck features with capsule networks to capture spatial relationships, and offers robustness against temporal shifts and AF perturbations. EfficientNet (*S2*) [30] applies compound scaling to balance detection accuracy and model efficiency to enable effective capture of localized spectral distortions. FFD (*S3*) [4] employs attention mechanisms to highlight manipulated regions within spectrogram images to detect deepfake traces. CORE [22] and FFD [4] generalize well to spectrogram-based ADD due to their architecture's focus on structural inconsistencies. CORE (*S4*) [22] introduces a consistency loss across augmented views to enforce invariant learning, thereby emphasizing intrinsic forgery cues over superficial artifacts. F3Net (*S5*) [26] integrates frequency decomposition with local frequency statistics to enhance detection performance. CQTNet (*S6*) [45] employs Constant-Q Transform spectrograms and self-attended ResNet blocks with one-class learning to improve generalization to unseen attacks.

2.3 Anti-Forensic Methods

We categorize the selected AF attacks into two primary classes: (a) statistical-based and (b) optimization-based methods. This taxonomy encompasses a representative range of AF strategies frequently observed in real-world deepfake generation and evasion scenarios.

2.3.1 Statistical AF Attacks

Statistical AF attacks, such as pitch shifting [3], filtering [41], noise addition [17], and compression [17], aim to alter the statistical properties, such as amplitude distribution, spectral content, or temporal patterns, without noticeably changing their perceptual quality. The primary goal is to evade ADD models that rely on these cues to identify synthetic nature.
Pitch Shifting [3] is a statistical AF method that alters amplitude distribution, spectral content, or temporal patterns without changing the audio's duration. It is often overlooked in deepfake detection, posing a potential vulnerability. We apply a phase vocoder for time-stretching, followed by resampling to shift the pitch as follows:

$$A' = \mathcal{F}^{-1}\left\{\mathcal{F}\left(A\right) \cdot H(f,n)\right\} \tag{1}$$

Here, F and F^{-1} denote the Fourier and inverse transforms of frequency f by n semitones. We apply pitch shifts of ±1, ±5, and ±12 semitones to simulate lowering and raising effects.
Filtering [41] is a technique used to remove or isolate specific frequency components of an audio signal. In ADD, median filtering can be used to simulate deepfake artifacts to create adversarial perturbations, defined as follows:

$$A' = \frac{A\left(\frac{N}{2}\right) + A\left(\frac{N}{2}+1\right)}{2} \tag{2}$$

We apply kernel sizes $N \in 3,5,7,9$ to control the strength of the AF perturbations.
Noise Addition [17] is a statistical approach that degrades the statistical patterns and spectral

features of audio to obscure deepfake artifacts. We applied Gaussian noise as a statistical AF attack, defined as follows:

$$A' = A + \delta; \quad \delta \sim \mathcal{N}(0, \sigma^2) \tag{3}$$

We set the standard deviation, σ, as 0.001, 0.01, 0.02, 0.03, 0.04, and 0.05.

Quantization [8] is a statistical transformation that reduces the precision of an audio by mapping its continuous amplitudes to a limited number of discrete levels. This process alters the fine-grained amplitude distribution and introduces quantization noise, thereby distorting the subtle statistical patterns that deepfake detectors often rely on, such as high-frequency details, spectral smoothness, and dynamic range. This is computed as follows:

$$A' = \frac{\text{round}\left((A+1) \times \left(\frac{L}{2} - 1\right)\right)}{\left(\frac{L}{2} - 1\right)} - 1 \tag{4}$$

where $L = 2^b$ is the quantization level. We selected the bit depth, b, as 4, 6, and 8 kbps.

2.3.2 Optimization-based AF Attacks

Optimization-based AFs, such as FGSM [7], PGD [19], C&W [2], and DeepFool [20], introduce minimal perturbations to fool ADDs while preserving perceptual quality. Unlike statistical methods, these attacks iteratively optimize a loss to maximize misclassification.

FGSM[7, 11] is a popular AF attack in computer vision and image processing that perturbs inputs in the direction of the gradient sign to induce misclassification. It effectively fools ADDs[44] by subtly altering audio without affecting perceptual quality. Let A be the input audio with label y. FGSM generates an attacked audio A' using the loss $J(\theta, A, y)$ as follows:

$$A' = A + \varepsilon \cdot \text{sign}\left(\nabla_A J(\theta, A, y)\right) \tag{5}$$

We set ε as 0.001, 0.05, 0.01, 0.1, and 0.2 to control the perturbation magnitude.

PGD [11, 19] is an iterative AF attack that applies small gradient-based perturbations at each step to project them within a bound to maintain imperceptibility. It produces stronger attacks than single-step methods, defined as:

$$A'_{t+1} = \begin{cases} A + \delta, & t = 0, \quad \delta \sim \mathcal{U}(-\varepsilon, \varepsilon) \\ \text{Proj}_\varepsilon\left(A'_t + \alpha \cdot \text{sign}\left(\nabla_A J(\theta, A'_t, y)\right)\right), & t \geq 0 \end{cases} \tag{6}$$

where A'_t and δ are the attacked sample at step t and random noise. We set ε as 0.003, 0.007, 0.015, 0.03, and 0.06 with a step size α of 20 to control the perturbation magnitude.

C & W [2] is a powerful and widely used AF attack that crafts a targeted sample with minimal distortion. It formulates the process as a constrained optimization problem to find the smallest perturbation, defined as:

$$\min_{\delta} \quad \|\delta\|_2^2 + c \cdot f(A+\delta), \quad \text{where} \quad f(A') = \max\left(\max_{i \neq y} Z_a(A') - Z_y(A'), -k\right) \tag{7}$$

where $Z_a(.)$ and $Z_y(.)$ indicate the logits of attack and ground truth. We set confidence c as 0, 10, 25, 35, and 50 to control perturbations and default k as 0.

DeepFool [20] is an optimization-based AF attack that iteratively perturbs inputs to cross the decision boundary with minimal distortion. Unlike FGSM and PGD, it computes the closest adversarial point using a linearized model approximation, defined as:

$$A' = A + \mathbf{r} = -\frac{Z_k(A) - Z_l(A)}{\|\nabla Z_k(A) - \nabla Z_l(A)\|_2^2} \cdot \left(\nabla Z_k(A) - \nabla Z_l(A)\right) \tag{8}$$

where $Z_k(A)$ and $Z_l(A)$ are the model's confidence for classes k and l. We set the step size to 50 and varied the overshoot in $0.005, 0.01, 0.02, 0.03, 0.05$ to balance the attack and quality.

3 Results

3.1 Experimental Setups

This section details the environmental, training, and testing settings for overall evaluation.

3.1.1 Environmental Configurations

All experiments were conducted on a Linux 24.04 system equipped with eight NVIDIA RTX 6000 Ada Generation GPUs, each with 49 GB of memory. We used Python 3.9 and PyTorch 2.2.2 as the software environment for implementing and evaluating the model.

3.1.2 Training and Testing Configurations

We combined datasets $D1$, $D2$, and $D4$ for training, encompassing diverse generative methods, compression formats, and codecs. Testing used official splits from all five datasets ($D1$–$D5$) to improve generalization. The ADD models are trained for 50 epochs with a learning rate of 0.0001, batch size 256, using the Adam optimizer with weight decay 0.0001.

3.2 Performance Evaluations and Comparisons

3.2.1 Baseline Detection Results

As discussed in Section 2.2, we evaluated two ADD groups: raw audio and spectrogram-based methods. Table 1 lists the baseline detection results. In particular, spectrogram-based methods ($S1$-$S6$, achieving an average AUC and EER of 0.86 and 0.19) work better than raw methods ($R1$-$R6$, achieving an average AUC and EER of 0.75 and 0.30). Since all ADD models are trained on $D1$, $D2$, and $D4$, they generalize well to unseen datasets $D3$ and $D5$.

3.3 Results against Statistical AF Attacks

We evaluated the resilience of ADDs against four statistical AFs: pitch shifting, filtering, noise addition, and quantization as described in Section 2.3.1. Table 2 summarizes the average AUC and EER scores in five datasets. For raw ADDs, performance degradation was substantial across all statistical AFs. Specifically, average AUC/EER dropped to 0.57/0.43 for pitch shifting, 0.59/0.42 for filtering, 0.52/0.47 for noise addition, and 0.60/0.42 for quantization. $R2$ [38], $R3$ [41], and $R4$ [11] were notably vulnerable to filtering and noise, with AUCs falling below 0.30. In contrast, $R1$ [10] and $R5$ [9] maintained relatively higher robustness due to enhanced temporal modeling and stronger feature representations.

On the other hand, spectrogram-based ADDs exhibited similar susceptibility to statistical AFs. As reported in Table 2, average AUC and EER were reduced to 0.50/0.44 for pitch shifting [3], 0.44/0.49 for filtering [41], 0.48/0.49 for noise addition [17], and 0.43/0.53 for quantization [8]. Despite strong visual learning, spectrogram-based methods show instability under simple AFs, emphasizing the need for more robust ADD methods.

3.4 Results against Optimization-based AF Attacks

We evaluated ADD robustness against four optimization-based AFs: FGSM, PGD, C&W, and DeepFool (Section 2.3.2). Table 3 shows that FGSM [7] dropped AUC to 0.35 and raised EER to 0.62, while PGD [19] further degraded performance (AUC: 0.09, EER: 0.87). C&W [2] and DeepFool [20] also caused notable declines (avg. AUCs: 0.51 and 0.55). $R2$, $R3$, and $R6$ were highly vulnerable, whereas $R1$ and $R5$ showed better robustness.

Table 1: Baseline Performance of ADD Methods.

Raw Audio-based ADD Methods							**Spectrogram-based ADD Methods**						
Method	D1 [38]	D2 [38]	D3 [38]	D4 [38]	D5 [0]	Avg.	Method	D1 [38]	D2 [38]	D3 [38]	D4 [38]	D5 [0]	Avg.
R1 [38]	0.99 / 0.05	0.73 / 0.32	0.68 / 0.36	0.83 / 0.25	0.96 / 0.10	0.83 / 0.21	*S1* [38]	0.99 / 0.04	0.78 / 0.27	0.57 / 0.45	0.93 / 0.13	0.89 / 0.19	0.83 / 0.22
R2 [38]	0.98 / 0.07	0.80 / 0.24	0.53 / 0.46	0.66 / 0.40	0.56 / 0.47	0.71 / 0.33	*S2* [38]	0.99 / 0.04	0.83 / 0.26	0.58 / 0.42	0.93 / 0.14	0.81 / 0.25	0.83 / 0.21
R3 [38]	0.93 / 0.16	0.78 / 0.28	0.58 / 0.45	0.87 / 0.23	0.54 / 0.49	0.74 / 0.32	*S3* [0]	1.00 / 0.04	0.85 / 0.25	0.64 / 0.40	0.99 / 0.06	0.99 / 0.04	0.89 / 0.16
R4 [38]	0.98 / 0.21	0.85 / 0.22	0.54 / 0.48	0.72 / 0.33	0.54 / 0.48	0.73 / 0.34	*S4* [38]	0.98 / 0.09	0.82 / 0.26	0.65 / 0.40	0.99 / 0.07	0.98 / 0.08	0.88 / 0.18
R5 [0]	0.98 / 0.06	0.75 / 0.31	0.65 / 0.39	0.86 / 0.23	0.67 / 0.38	0.78 / 0.27	*S5* [38]	0.99 / 0.04	0.79 / 0.29	0.63 / 0.40	0.97 / 0.08	0.96 / 0.11	0.87 / 0.18
R6 [0]	0.99 / 0.05	0.76 / 0.30	0.52 / 0.48	0.75 / 0.34	0.61 / 0.42	0.73 / 0.32	*S6* [38]	0.98 / 0.06	0.70 / 0.36	0.66 / 0.40	0.97 / 0.10	0.98 / 0.07	0.86 / 0.20
Avg.	**0.98 / 0.10**	**0.78 / 0.28**	**0.58 / 0.43**	**0.78 / 0.30**	**0.65 / 0.39**	**0.75 / 0.30**	**Avg.**	**0.99 / 0.05**	**0.80 / 0.28**	**0.62 / 0.41**	**0.96 / 0.10**	**0.94 / 0.12**	**0.86 / 0.19**

Table 2: Performance of Statistical AF Attacks on ADD Methods.

Raw Audio-based ADD Methods							Spectrogram-based ADD Methods						
Method	D1	D2	D3	D4	D5	Avg.	Method	D1	D2	D3	D4	D5	Avg.
Pitch Shifting							**Pitch Shifting**						
R1	0.92 / .12	0.50 / 0.51	0.43 / 0.56	0.75 / 0.30	0.75 / 0.30	0.67 / 0.36	*S1*	0.31 / 0.84	0.54 / 0.23	0.68 / 0.32	0.31 / 0.67	0.56 / 0.23	0.48 / 0.46
R2	0.86 / 0.21	0.56 / 0.42	0.30 / 0.65	0.54 / 0.44	0.29 / 0.63	0.51 / 0.47	*S2*	0.37 / 0.63	0.40 / 0.60	0.37 / 0.63	0.56 / 0.44	0.41 / 0.58	0.42 / 0.58
R3	0.92 / 0.13	0.60 / 0.41	0.19 / 0.74	0.80 / 0.27	0.15 / 0.78	0.53 / 0.47	*S3*	0.35 / 0.83	0.51 / 0.25	0.66 / 0.35	0.59 / 0.39	0.32 / 0.84	0.49 / 0.53
R4	0.80 / 0.25	0.81 / 0.23	0.36 / 0.60	0.51 / 0.50	0.12 / 0.77	0.52 / 0.47	*S4*	0.47 / 0.76	0.44 / 0.78	0.69 / 0.30	0.59 / 0.41	0.37 / 0.80	0.51 / 0.61
R5	0.95 / 0.12	0.70 / 0.36	0.30 / 0.65	0.84 / 0.22	0.39 / 0.56	0.64 / 0.38	*S5*	0.58 / 0.21	0.53 / 0.24	0.54 / 0.46	0.33 / 0.66	0.53 / 0.25	0.50 / 0.36
R6	0.93 / 0.14	0.41 / 0.52	0.23 / 0.69	0.78 / 0.28	0.33 / 0.61	0.54 / 0.45	*S6*	0.54 / 0.22	0.50 / 0.26	0.96 / 0.11	0.34 / 0.68	0.51 / 0.26	0.57 / 0.31
Avg.	**0.90 / 0.16**	**0.60 / 0.41**	**0.30 / 0.65**	**0.70 / 0.34**	**0.34 / 0.61**	**0.57 / 0.43**	**Avg.**	**0.45 / 0.45**	**0.49 / 0.39**	**0.64 / 0.36**	**0.45 / 0.54**	**0.45 / 0.49**	**0.50 / 0.44**
Filtering							**Filtering**						
R1	0.98 / 0.03	0.88 / 0.15	0.73 / 0.28	0.96 / 0.10	0.98 / 0.04	0.91 / 0.12	*S1*	0.38 / 0.81	0.58 / 0.21	0.30 / 0.70	0.35 / 0.65	0.43 / 0.78	0.41 / 0.63
R2	0.92 / 0.15	0.48 / 0.49	0.21 / 0.73	0.82 / 0.25	0.52 / 0.49	0.59 / 0.42	*S2*	0.35 / 0.65	0.46 / 0.53	0.47 / 0.53	0.31 / 0.69	0.53 / 0.45	0.42 / 0.57
R3	0.34 / 0.60	0.14 / 0.75	0.03 / 0.90	0.52 / 0.47	0.14 / 0.78	0.23 / 0.70	*S3*	0.35 / 0.83	0.49 / 0.26	0.27 / 0.74	0.36 / 0.65	0.31 / 0.84	0.36 / 0.66
R4	0.35 / 0.62	0.11 / 0.82	0.03 / 0.94	0.10 / 0.83	0.03 / 0.97	0.12 / 0.84	*S4*	0.60 / 0.20	0.31 / 0.84	0.35 / 0.65	0.47 / 0.52	0.57 / 0.23	0.46 / 0.49
R5	0.97 / 0.08	0.81 / 0.24	0.51 / 0.48	0.95 / 0.11	0.86 / 0.21	0.82 / 0.22	*S5*	0.60 / 0.20	0.58 / 0.22	0.40 / 0.59	0.47 / 0.53	0.56 / 0.23	0.52 / 0.35
R6	0.99 / 0.01	0.77 / 0.25	0.69 / 0.35	0.98 / 0.06	0.76 / 0.33	0.84 / 0.20	*S6*	0.56 / 0.23	0.54 / 0.24	0.30 / 0.69	0.38 / 0.62	0.52 / 0.24	0.46 / 0.40
Avg.	**0.76 / 0.25**	**0.53 / 0.45**	**0.37 / 0.61**	**0.72 / 0.30**	**0.55 / 0.47**	**0.59 / 0.42**	**Avg.**	**0.47 / 0.35**	**0.48 / 0.38**	**0.35 / 0.65**	**0.39 / 0.61**	**0.49 / 0.46**	**0.44 / 0.49**
Noise Addition							**Noise Addition**						
R1	0.97 / 0.04	0.71 / 0.32	0.46 / 0.55	0.94 / 0.10	0.97 / 0.07	0.81 / 0.22	*S1*	0.49 / 0.75	0.41 / 0.79	0.48 / 0.53	0.52 / 0.49	0.38 / 0.81	0.46 / 0.67
R2	0.64 / 0.40	0.32 / 0.62	0.25 / 0.66	0.10 / 0.88	0.13 / 0.79	0.29 / 0.67	*S2*	0.37 / 0.63	0.51 / 0.50	0.34 / 0.66	0.54 / 0.46	0.41 / 0.58	0.43 / 0.57
R3	0.74 / 0.32	0.37 / 0.60	0.14 / 0.80	0.76 / 0.31	0.20 / 0.74	0.44 / 0.55	*S3*	0.46 / 0.77	0.48 / 0.76	0.21 / 0.80	0.54 / 0.46	0.34 / 0.83	0.41 / 0.72
R4	0.73 / 0.28	0.11 / 0.82	0.09 / 0.85	0.23 / 0.71	0.04 / 0.87	0.24 / 0.71	*S4*	0.57 / 0.22	0.59 / 0.21	0.67 / 0.32	0.54 / 0.47	0.50 / 0.26	0.57 / 0.30
R5	0.89 / 0.17	0.81 / 0.24	0.19 / 0.76	0.92 / 0.16	0.66 / 0.40	0.69 / 0.35	*S5*	0.37 / 0.81	0.53 / 0.24	0.59 / 0.41	0.39 / 0.61	0.55 / 0.23	0.49 / 0.46
R6	0.95 / 0.10	0.77 / 0.25	0.24 / 0.70	0.85 / 0.24	0.47 / 0.52	0.66 / 0.36	*S6*	0.55 / 0.24	0.52 / 0.25	0.54 / 0.46	0.45 / 0.55	0.53 / 0.23	0.52 / 0.35
Avg.	**0.77 / 0.22**	**0.51 / 0.47**	**0.23 / 0.71**	**0.63 / 0.40**	**0.41 / 0.56**	**0.52 / 0.47**	**Avg.**	**0.46 / 0.44**	**0.52 / 0.46**	**0.46 / 0.53**	**0.50 / 0.51**	**0.45 / 0.49**	**0.48 / 0.49**
Quantization							**Quantization**						
R1	0.88 / 0.20	0.56 / 0.47	0.46 / 0.49	0.68 / 0.39	0.64 / 0.39	0.64 / 0.39	*S1*	0.37 / 0.82	0.47 / 0.76	0.24 / 0.76	0.52 / 0.48	0.31 / 0.84	0.38 / 0.73
R2	0.92 / 0.23	0.49 / 0.46	0.30 / 0.63	0.59 / 0.42	0.34 / 0.60	0.53 / 0.47	*S2*	0.49 / 0.51	0.41 / 0.60	0.52 / 0.48	0.34 / 0.66	0.45 / 0.53	0.44 / 0.56
R3	0.85 / 0.23	0.66 / 0.36	0.28 / 0.66	0.40 / 0.58	0.04 / 0.89	0.45 / 0.54	*S3*	0.56 / 0.22	0.34 / 0.83	0.50 / 0.50	0.28 / 0.73	0.48 / 0.76	0.43 / 0.61
R4	0.83 / 0.22	0.68 / 0.33	0.14 / 0.79	0.53 / 0.48	0.06 / 0.85	0.45 / 0.53	*S4*	0.56 / 0.22	0.60 / 0.20	0.47 / 0.51	0.42 / 0.57	0.35 / 0.81	0.48 / 0.46
R5	0.99 / 0.03	0.79 / 0.28	0.69 / 0.35	0.85 / 0.23	0.69 / 0.37	0.80 / 0.25	*S5*	0.35 / 0.83	0.39 / 0.80	0.44 / 0.57	0.31 / 0.69	0.38 / 0.80	0.37 / 0.74
R6	0.99 / 0.04	0.74 / 0.30	0.49 / 0.51	0.82 / 0.28	0.54 / 0.47	0.72 / 0.32	*S6*	0.53 / 0.25	0.56 / 0.22	0.33 / 0.67	0.36 / 0.64	0.49 / 0.25	0.45 / 0.40
Avg.	**0.91 / 0.15**	**0.65 / 0.37**	**0.39 / 0.57**	**0.64 / 0.40**	**0.38 / 0.59**	**0.60 / 0.42**	**Avg.**	**0.48 / 0.44**	**0.46 / 0.57**	**0.42 / 0.57**	**0.37 / 0.62**	**0.41 / 0.66**	**0.43 / 0.53**

Spectrogram-based ADDs were similarly impacted by optimization-based AFs (Table 3). FGSM and PGD dropped AUC to 0.35 and raised EER above 0.60. C & W and DeepFool further degraded performance. While *S*1 and *S*4 showed slight robustness due to attention, *S*2 and *S*5 were moderately affected. These results reveal the need for robust models.

3.5 Defense against AF Attacks

To counter the AFs described in Section 2.3, we selected four top-performing ADDs: two raw-based (*R*1, *R*5) and two spectrogram-based (*S*3, *S*4) methods. These were trained with adversarial examples generated using random parameter selections. Table 4 shows the detection AUC and EER for both seen and unseen datasets. Compared to the results in Tables 2 and 3, adversarial training improves the average AUC and EER to 0.76 and 0.28 for raw ADDs, and 0.83 and 0.18 for spectrogram-based ADDs. However, these values remain lower than the baseline ADD results, indicating the need for more robust ADDs to defend against AFs.

3.6 Empirical Analysis

To assess audio quality, we computed average mean square error (MSE) and structural similarity index measure (SSIM) between original and attacked audios (Table 5). For statistical AFs, we used: pitch shift = –1 semitone, median filter size = 3, noise std = 0.001, and bit depth = 4. For optimization-based AFs, parameters were: FGSM $\varepsilon = 0.001$, PGD $\varepsilon = 0.003$, C & W confidence = 10, DeepFool overshoot = 0.005, with *R*4 as the detection model.

As shown in Table 5, noise and quantization cause less distortion than pitch shifting and filtering, which alter the audio's structure more noticeably. Still, MSE stays below 0.01 and SSIM above 0.90, indicating minimal overall distortion. In contrast, optimization-based AFs introduce significantly less distortion compared to statistical AFs. The average MSE remains below 0.009 and SSIM above 0.95, which indicates high similarity to the input samples. Compared to statistical AF attacks, optimization-based methods generate AF attacks that preserve the audio quality better.

Table 3: Performance of Optimization-based AF Attacks on ADD Methods.

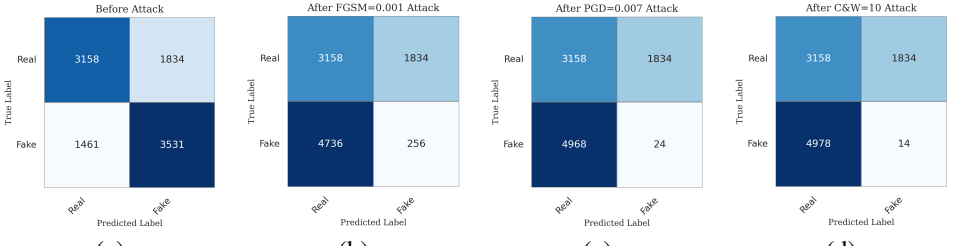
Raw Audio-based ADD Methods							**Spectrogram-based ADD Methods**						
Method	**D1**	**D2**	**D3**	**D4**	**D5**	**Avg.**	Method	**D1**	**D2**	**D3**	**D4**	**D5**	**Avg.**
FGSM							**FGSM**						
R1	0.69 / 0.37	0.33 / 0.65	0.17 / 0.78	0.69 / 0.36	0.77 / 0.31	0.53 / 0.50	*S1*	0.52 / 0.24	0.40 / 0.80	0.21 / 0.85	0.55 / 0.44	0.30 / 0.62	0.40 / 0.59
R2	0.21 / 0.73	0.03 / 0.94	0.01 / 0.97	0.02 / 0.97	0.00 / 1.00	0.05/0.92	*S2*	0.47 / 0.52	0.49 / 0.50	0.23 / 0.63	0.54 / 0.46	0.44 / 0.53	0.43 / 0.53
R3	0.69 / 0.39	0.24 / 0.67	0.21 / 0.68	0.64 / 0.38	0.17 / 0.82	0.39 / 0.59	*S3*	0.35 / 0.82	0.22 / 0.89	0.21 / 0.74	0.42 / 0.58	0.27 / 0.86	0.29 / 0.78
R4	0.59 / 0.38	0.30 / 0.59	0.13 / 0.75	0.32 / 0.59	0.09 / 0.78	0.28 / 0.62	*S4*	0.33 / 0.83	0.33 / 0.53	0.25 / 0.50	0.47 / 0.57	0.22 / 0.89	0.32 / 0.67
R5	0.88 / 0.20	0.58 / 0.44	0.20 / 0.72	0.91 / 0.17	0.51 / 0.48	0.62 / 0.40	*S5*	0.31 / 0.85	0.52 / 0.25	0.25 / 0.50	0.26 / 0.74	0.53 / 0.24	0.37 / 0.52
R6	0.70 / 0.34	0.16 / 0.75	0.01 / 0.94	0.29 / 0.64	0.05 / 0.87	0.24 / 0.71	*S6*	0.40 / 0.65	0.39 / 0.60	0.23 / 0.64	0.45 / 0.56	0.35 / 0.63	0.37 / 0.62
Avg.	**0.63 / 0.40**	**0.27 / 0.67**	**0.12 / 0.81**	**0.48 / 0.52**	**0.26 / 0.71**	**0.35 / 0.62**	**Avg.**	**0.40 / 0.65**	**0.39 / 0.59**	**0.23 / 0.64**	**0.45 / 0.56**	**0.35 / 0.63**	**0.36 / 0.61**
PGD							**PGD**						
R1	0.41 / 0.60	0.25 / 0.69	0.12 / 0.82	0.23 / 0.71	0.43 / 0.57	0.29 / 0.68	*S1*	0.51 / 0.24	0.34 / 0.83	0.21 / 0.85	0.41 / 0.58	0.34 / 0.60	0.36 / 0.62
R2	0.09 / 0.87	0.01 / 0.98	0.10 / 0.90	0.04 / 0.96	0.03 / 0.97	0.05 / 0.94	*S2*	0.35 / 0.65	0.43 / 0.54	0.23 / 0.63	0.33 / 0.67	0.27 / 0.74	0.32 / 0.65
R3	0.03 / 0.97	0.02 / 0.93	0.05 / 0.95	0.03 / 0.93	0.07 / 0.93	0.04 / 0.94	*S3*	0.57 / 0.22	0.24 / 0.88	0.21 / 0.74	0.28 / 0.72	0.54 / 0.23	0.37 / 0.56
R4	0.20 / 0.68	0.10 / 0.80	0.04 / 0.96	0.02 / 0.95	0.09 / 0.91	0.09 / 0.86	*S4*	0.34 / 0.83	0.27 / 0.56	0.25 / 0.50	0.56 / 0.44	0.37 / 0.81	0.36 / 0.63
R5	0.07 / 0.85	0.02 / 0.91	0.11 / 0.89	0.02 / 0.93	0.04 / 0.96	0.05 / 0.91	*S5*	0.40 / 0.80	0.30 / 0.84	0.25 / 0.50	0.23 / 0.76	0.52 / 0.25	0.34 / 0.63
R6	0.10 / 0.82	0.01 / 0.98	0.09 / 0.91	0.01 / 0.99	0.05 / 0.95	0.05 / 0.93	*S6*	0.43 / 0.55	0.32 / 0.73	0.23 / 0.64	0.36 / 0.63	0.41 / 0.53	0.35 / 0.62
Avg.	**0.15 / 0.80**	**0.07 / 0.88**	**0.08 / 0.91**	**0.06 / 0.91**	**0.12 / 0.88**	**0.09 / 0.87**	**Avg.**	**0.43 / 0.55**	**0.32 / 0.73**	**0.23 / 0.64**	**0.36 / 0.63**	**0.41 / 0.53**	**0.35 / 0.62**
C & W							**C & W**						
R1	0.98 / 0.08	0.81 / 0.23	0.55 / 0.41	0.62 / 0.43	0.94 / 0.11	0.78 / 0.25	*S1*	0.57 / 0.21	0.29 / 0.85	0.21 / 0.85	0.53 / 0.47	0.38 / 0.59	0.40 / 0.59
R2	0.85 / 0.23	0.26 / 0.66	0.17 / 0.78	0.44 / 0.54	0.33 / 0.61	0.41 / 0.56	*S2*	0.23 / 0.77	0.55 / 0.43	0.23 / 0.63	0.33 / 0.67	0.21 / 0.78	0.31 / 0.66
R3	0.93 / 0.16	0.78 / 0.29	0.49 / 0.49	0.71 / 0.35	0.30 / 0.68	0.64 / 0.39	*S3*	0.50 / 0.25	0.36 / 0.81	0.21 / 0.74	0.41 / 0.59	0.50 / 0.26	0.40 / 0.53
R4	0.81 / 0.23	0.73 / 0.28	0.23 / 0.74	0.42 / 0.54	0.05 / 0.87	0.45 / 0.53	*S4*	0.20 / 0.90	0.50 / 0.44	0.25 / 0.50	0.43 / 0.59	0.47 / 0.76	0.37 / 0.64
R5	0.85 / 0.23	0.51 / 0.49	0.12 / 0.81	0.63 / 0.41	0.24 / 0.70	0.47 / 0.51	*S5*	0.58 / 0.21	0.33 / 0.82	0.25 / 0.50	0.40 / 0.62	0.27 / 0.86	0.37 / 0.60
R6	0.82 / 0.25	0.27 / 0.66	0.02 / 0.95	0.18 / 0.75	0.06 / 0.86	0.27 / 0.69	*S6*	0.42 / 0.47	0.41 / 0.67	0.23 / 0.64	0.42 / 0.59	0.37 / 0.65	0.37 / 0.60
Avg	**0.87 / 0.20**	**0.56 / 0.43**	**0.26 / 0.70**	**0.51 / 0.49**	**0.32 / 0.64**	**0.51 / 0.47**	**Avg.**	**0.42 / 0.47**	**0.41 / 0.67**	**0.23 / 0.64**	**0.42 / 0.59**	**0.37 / 0.65**	**0.37 / 0.60**
DeepFool							**DeepFool**						
R1	0.98 / 0.07	0.70 / 0.33	0.55 / 0.41	0.76 / 0.30	0.96 / 0.11	0.79 / 0.24	*S1*	0.56 / 0.22	0.26 / 0.87	0.21 / 0.85	0.41 / 0.58	0.27 / 0.64	0.34 / 0.63
R2	0.96 / 0.12	0.63 / 0.31	0.21 / 0.74	0.62 / 0.42	0.48 / 0.50	0.57 / 0.40	*S2*	0.22 / 0.78	0.51 / 0.46	0.23 / 0.63	0.53 / 0.47	0.21 / 0.78	0.34 / 0.62
R3	0.91 / 0.17	0.54 / 0.43	0.14 / 0.83	0.69 / 0.37	0.25 / 0.66	0.50 / 0.39	*S3*	0.38 / 0.81	0.51 / 0.25	0.21 / 0.74	0.26 / 0.74	0.55 / 0.23	0.38 / 0.55
R4	0.79 / 0.24	0.72 / 0.28	0.23 / 0.71	0.37 / 0.56	0.12 / 0.77	0.41 / 0.49	*S4*	0.58 / 0.21	0.54 / 0.42	0.25 / 0.50	0.41 / 0.60	0.32 / 0.84	0.42 / 0.51
R5	0.90 / 0.18	0.57 / 0.44	0.15 / 0.76	0.80 / 0.26	0.38 / 0.58	0.64 / 0.34	*S5*	0.30 / 0.85	0.48 / 0.28	0.25 / 0.50	0.37 / 0.63	0.28 / 0.85	0.34 / 0.62
R6	0.88 / 0.19	0.33 / 0.53	0.03 / 0.92	0.44 / 0.52	0.10 / 0.78	0.44 / 0.46	*S6*	0.41 / 0.57	0.46 / 0.38	0.23 / 0.64	0.40 / 0.60	0.33 / 0.67	0.37 / 0.57
Avg	**0.90 / 0.16**	**0.68 / 0.34**	**0.24 / 0.70**	**0.61 / 0.35**	**0.38 / 0.37**	**0.55 / 0.47**	**Avg.**	**0.41 / 0.57**	**0.46 / 0.38**	**0.23 / 0.64**	**0.40 / 0.60**	**0.33 / 0.67**	**0.37 / 0.57**

Table 4: Performance of Defense against AF Attacks.

Raw Audio-based ADD Methods							**Spectrogram-based ADD Methods**						
Method	**D1**	**D2**	**D3**	**D4**	**D5**	**Avg.**	Method	**D1**	**D2**	**D3**	**D4**	**D5**	**Avg.**
R1	0.95 / 0.06	0.54 / 0.41	0.74 / 0.35	0.91 / 0.10	0.87 / 0.19	0.80 / 0.22	*S3*	0.99 / 0.06	0.83 / 0.11	0.61 / 0.42	0.99 / 0.01	0.82 / 0.25	0.85 / 0.17
R5	0.89 / 0.13	0.67 / 0.38	0.58 / 0.43	0.78 / 0.31	0.61 / 0.44	0.71 / 0.34	*S4*	0.98 / 0.07	0.79 / 0.12	0.59 / 0.45	0.99 / 0.01	0.75 / 0.30	0.82 / 0.19
Avg	**0.92 / 0.10**	**0.61 / 0.40**	**0.66 / 0.39**	**0.85 / 0.21**	**0.74 / 0.32**	**0.76 / 0.28**	**Avg**	**0.99 / 0.06**	**0.81 / 0.12**	**0.60 / 0.44**	**0.99 / 0.01**	**0.79 / 0.28**	**0.83 / 0.18**

Table 5: Qualitative Results of non-Attacked and Attacked Samples.

Raw Audio-based ADD Methods							**Spectrogram-based ADD Methods**							
Method	**Metric**	**D1**	**D2**	**D3**	**D4**	**D5**	**Method**	**Metric**	**D1**	**D2**	**D3**	**D4**	**D5**
Pitch Shifting	MSE	0.0334	0.0218	0.005	0.0005	0.0038	FGSM	MSE	0.0095	0.0098	0.0091	0.0046	0.0017
	SSIM	0.887	0.872	0.954	0.947	0.919		SSIM	0.946	0.95	0.969	0.957	0.952
Filtering	MSE	0.0235	0.0146	0.0032	0.0004	0.0025	PGD	MSE	0.0095	0.0098	0.0091	0.0046	0.0017
	SSIM	0.819	0.776	0.927	0.911	0.857		SSIM	0.947	0.952	0.969	0.958	0.953
Noise Addition	MSE	0.0002	0.0006	0.0001	0.0002	0.0001	C & W	MSE	0.0094	0.0098	0.009	0.0042	0.0017
	SSIM	0.991	0.989	0.994	0.988	0.990		SSIM	0.947	0.95	0.97	0.96	0.952
Quantization	MSE	0.001	0.0011	0.0006	0.0003	0.0008	DeepFool	MSE	0.0069	0.0069	0.0072	0.0022	0.0012
	SSIM	0.999	0.994	0.996	0.922	0.968		SSIM	0.96	0.962	0.976	0.971	0.968
Avg MSE		**0.0145**	**0.0095**	**0.0022**	**0.0004**	**0.0018**	**Avg MSE**		**0.0088**	**0.0091**	**0.0086**	**0.0039**	**0.0016**
Avg SSIM		**0.924**	**0.908**	**0.968**	**0.942**	**0.933**	**Avg SSIM**		**0.950**	**0.954**	**0.971**	**0.962**	**0.956**



(a) (b) (c) (d)

Figure 2: Confusion matrix before and after AF attacks; (a) Original, (b) FGSM, (c) PGD, and (d) C & W using *R*4 on the *D*4 dataset.

4 Discussions and Future Directions

4.1 Analysis and Discussions

This study provides a comprehensive evaluation of SoTA ADDs under various AF scenarios. The proposed comparative analysis highlights that while current raw and spectrogram-based ADDs perform reasonably well on clean data (randomly selected 10000 samples from $D5$), as shown in Figure 2 (a), their effectiveness significantly degrades under AFs as depicted in Figure 2 (b)-(d). Statistical AFs, such as pitch shifting, filtering, noise addition, and quantization, can obscure generative AI signatures as visualized in Figure 3 (a) & (b). In contrast, optimization-based AFs (e.g., FGSM, DeepFool) expose fundamental vulnerabilities of ADDs by exploiting their targeted perturbations, as shown in Figure 3 (c) & (d).

The performance drop, as shown in Figures 3 (a)-(d), underscores the critical challenge in designing resilient ADDs that can adapt to diverse AFs. Furthermore, the qualitative analysis, including spectrogram visualization as given in Figure 4 (a) & (b), reveals that some AFs introduce subtle distortions 4 (a) for pitch shifting that may go unnoticed by human listeners but are sufficient to deceive ADDs. These suggest that while spectrogram-based methods benefit from well-established backbone models, raw audio-based models retain advantages in capturing temporal dynamics, indicating that a hybrid approach could be beneficial.

Overall, these findings underscore the need to move beyond conventional supervised learning toward adaptive and explainable models that can effectively mitigate sophisticated AFs.

4.2 Challenges and Future Directions

Evolving Nature of AF Attacks: This comparative analysis offers valuable insights into how different AF attacks affect ADD methods and helps researchers better understand AF behaviors and identify blind spots of detectors.

Model Vulnerability: A key challenge revealed by our study is that current ADDs exhibit architecture-specific vulnerabilities. This calls for future research into designing more resilient architectures and learning strategies that can adapt to diverse AF techniques.

Standardized Evaluation Framework: This study introduces a unified and reproducible evaluation pipeline by leveraging a wide range of ADD and AF techniques, providing a consistent foundation for future research and comparison within the ADD community.

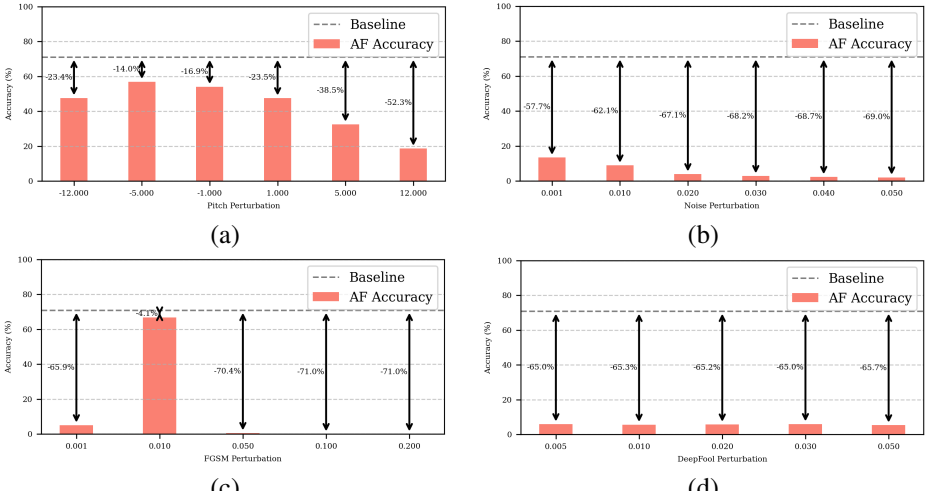
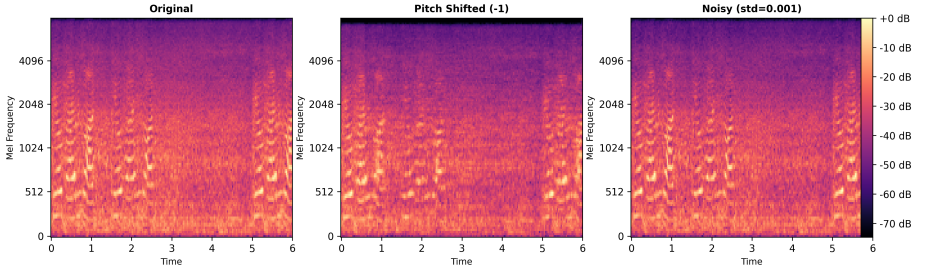
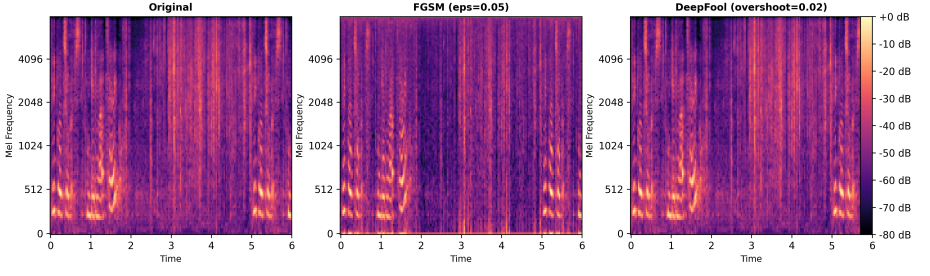


Figure 3: Accuracy of AFs on ADD methods; (a) Pitch Shifting, (b) Noise Addition, (c) FGSM, and (d) DeepFool with $R4$ on the $D4$ dataset.



(a)



(b)

Figure 4: Spectrogram visualization before and after AF attacks; (a) Original, Pitch, and Noise, (b) Original, FGSM, and DeepFool.

Foundation for Robust Model Design: The performance drops under AF attacks guides researchers toward designing advanced techniques like adversarial training, data augmentation, and domain adaptation to enable more robust and generalized systems.
Multimodal Expansion: The findings encourage future studies to incorporate complementary multimodal (e.g., video, lip-sync) to enhance robustness in real-world applications.

5 Conclusions

Nowadays, deepfake audio is posing substantial risks to voice biometrics and related applications. While several SoTA ADDs, including raw and spectrogram-based approaches, demonstrate promising performance in identifying generative AI artifacts, their effectiveness is critically challenged by diverse AFs. These attacks, particularly statistical (pitch shifting, filtering, noise addition, and quantization) and optimization-based methods (FGSM, PGD, C & W, and DeepFool), effectively alter or conceal AI signatures. The proposed comparative analysis across benchmark datasets highlights the vulnerabilities of both raw and spectrogram-based ADDs under AFs. These findings emphasize the urgent need for more resilient and generalized ADDs capable of countering diverse AF techniques.
In future, we aim to design adaptive and explainable study that cover a wider range of forensics and AFs, including foundation models and generative AFs. Exploring multi-modal approaches that combine audio and visual modalities will further enhance research directions. Additionally, integrating adversarial training and meta-learning techniques holds promise for improving generalization and robustness against unseen AF attack types.

Acknowledgments

This material is based upon work supported by the National Science Foundation (NSF) under Grant number 2409577. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

References

[1] Shaima M Alghamdi, Salma Kammoun Jarraya, and Faris Kateb. Enhancing security in multimodal biometric fusion: Analyzing adversarial attacks. *IEEE Access*, 2024.

[2] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

[3] Artem Chirkovskiy, Marina Volkova, and Nikita Khmelev. Pitch-shifted speech detection using audio ssl model. In *2025 27th International Conference on Digital Signal Processing and its Applications (DSPA)*, pages 1–7. IEEE, 2025.

[4] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*, pages 5781–5790, 2020.

[5] Muhammad Umar Farooq, Awais Khan, Kutub Uddin, and Khalid Mahmood Malik. Transferable adversarial attacks on audio deepfake detection. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 1640–1649, 2025.

[6] Joel Frank and Lea Schönherr. Wavefake: A data set to facilitate audio deepfake detection. *arXiv preprint arXiv:2111.02813*, 2021.

[7] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[8] Lăcrimioara Grama, Corneliu Rusu, Gabriel Oltean, and Laura Ivanciu. Quantization effects on audio signals for detecting intruders in wild areas using tespar s-matrix and artificial neural networks. In *2015 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–6. IEEE, 2015.

[9] Guang Hua, Andrew Beng Jin Teoh, and Haijian Zhang. Towards end-to-end synthetic speech detection. *IEEE Signal Processing Letters*, 28:1265–1269, 2021.

[10] Jee-weon Jung, You Jin Kim, Hee-Soo Heo, Bong-Jin Lee, Youngki Kwon, and Joon Son Chung. Pushing the limits of raw waveform speaker recognition. *arXiv preprint arXiv:2203.08488*, 2022.

[11] Patryk Kawa, Mateusz Plata, and Paweł Syga. Defense against adversarial attacks on audio deepfake detection. In *Proceedings of Interspeech 2023*, pages 5276–5280, 2023. doi: 10.21437/Interspeech.2023-409.

[12] Awais Khan and Khalid Mahmood Malik. Spotnet: A spoofing-aware transformer network for effective synthetic speech detection. In *Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation*, MAD '23, page 10–18, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701870. doi: 10.1145/3592572.3592841. URL https://doi.org/10.1145/3592572.3592841.

[13] Awais Khan and Khalid Mahmood Malik. Securing voice biometrics: One-shot learning approach for audio deepfake detection. In *2023 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, 2023. doi: 10.1109/WIFS58808.2023.10374968.

[14] Awais Khan, Khalid Mahmood Malik, James Ryan, and Mikul Saravanan. Battling voice spoofing: a review, comparative analysis, and generalizability evaluation of state-of-the-art voice spoofing counter measures. *Artif. Intell. Rev.*, 56(Suppl 1):513–566, June 2023. ISSN 0269-2821. doi: 10.1007/s10462-023-10539-8. URL https://doi.org/10.1007/s10462-023-10539-8.

[15] Awais Khan, Ijaz Ul Haq, and Khalid Mahmood Malik. Parallel stacked aggregated network for voice authentication in iot-enabled smart devices, 2024. URL https://arxiv.org/abs/2411.19841.

[16] Awais Khan, Khalid Mahmood Malik, and Shah Nawaz. Frame-to-utterance convergence: A spectra-temporal approach for unified spoofing detection. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10761–10765, 2024. doi: 10.1109/ICASSP48485.2024.10447500.

[17] Xiang Li, Pin-Yu Chen, and Wenqi Wei. Measuring the robustness of audio deepfake detectors. *arXiv preprint arXiv:2503.17577*, 2025.

[18] Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, et al. Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2507–2522, 2023.

[19] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[20] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.

[21] Paniti Netinant, Thitipong Utsanok, Meennapa Rukhiran, and Suttipong Klongdee. Development and assessment of internet of things-driven smart home security and automation with voice commands. *IoT*, 5(1):79–99, 2024.

[22] Yunsheng Ni, Depu Meng, Changqian Yu, Chengbin Quan, Dongchun Ren, and Youjian Zhao. Core: Consistent representation learning for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12–21, 2022.

[23] Seong Hee Park, Soo-Hyun Lee, Min Young Lim, Pyo Min Hong, and Youn Kyu Lee. A comprehensive risk analysis method for adversarial attacks on biometric authentication systems. *IEEE Access*, 2024.

[24] Pindrop. 2025 voice intelligence & security report. https://www.pindrop.com/research/report/voice-intelligence-security-report/, 2023. Accessed: 2025-07-11.

[25] Pindrop. Pindrop's 2025 Voice Intelligence & Security Report Reveals 1,300% Surge in Deepfake Fraud. https://www.prnewswire.com/news-releases/pindrops-2025-voice-intelligence--security-report-reveals-1-3

html, 2024. URL https://www.prnewswire.com/news-releases/pindrops-2025-voice-intelligence--security-report-reveals-1-html. Accessed: 2025-07-11.

[26] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pages 86–103. Springer, 2020.

[27] Mouna Rabhi, Spiridon Bakiras, and Roberto Di Pietro. Audio-deepfake detection: Adversarial attacks and countermeasures. *Expert Systems with Applications*, 250:123941, 2024.

[28] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. End-to-end anti-spoofing with rawnet2. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6369–6373. IEEE, 2021.

[29] Hemlata Tak, Madhu Kamble, Jose Patino, Massimiliano Todisco, and Nicholas Evans. Rawboost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6382–6386. IEEE, 2022.

[30] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[31] Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee. Asvspoof 2019: Future horizons in spoofed and fake audio detection. *arXiv preprint arXiv:1904.05441*, 2019.

[32] Kutub Uddin and Byung Tae Oh. Enhanced adversarial attack for avoidance of fake image detection. *Journal of Broadcast Engineering*, 28(7):859–866.

[33] Kutub Uddin, Yoonmo Yang, and Byung Tae Oh. Anti-forensic against double jpeg compression detection using adversarial generative network. In *Proceedings of the Korean Society of Broadcast Engineers Conference*, pages 58–60. The Korean Institute of Broadcast and Media Engineers, 2019.

[34] Kutub Uddin, Yoonmo Yang, Tae Hyun Jeong, and Byung Tae Oh. A robust open-set multi-instance learning for defending adversarial attacks in digital image. *IEEE Transactions on Information Forensics and Security*, 19:2098–2111, 2023.

[35] Kutub Uddin, Yoonmo Yang, and Byung Tae Oh. Deep learning-based counter anti-forensic of gan-based attack in hevc compressed domain using coding pattern analysis. *Expert Systems with Applications*, 233:120912, 2023.

[36] Kutub Uddin, Tae Hyun Jeong, and Byung Tae Oh. Counter-act against gan-based attacks: A collaborative learning approach for anti-forensic detection. *Applied Soft Computing*, 153:111287, 2024.

[37] Kutub Uddin, Awais Khan, Muhammad Umar Farooq, and Khalid Malik. Shield: A secure and highly enhanced integrated learning for robust deepfake detection against adversarial attacks, 2025. URL https://arxiv.org/abs/2507.13170. arXiv preprint arXiv:2507.13170.

[38] Fei Wang, Jinsong Han, Shiyuan Zhang, Xu He, and Dong Huang. Csi-net: Unified human body characterization and pose recognition. *arXiv preprint arXiv:1810.03064*, 2018.

[39] Xin Wang, Héctor Delgado, Hemlata Tak, Jee-weon Jung, Hye-jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi Kinnunen, et al. Asvspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale. *arXiv preprint arXiv:2408.08739*, 2024.

[40] Taiba Majid Wani, Reeva Gulzar, and Irene Amerini. Abc-capsnet: Attention based cascaded capsule network for audio deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2464–2472, 2024.

[41] Haibin Wu, Songxiang Liu, Helen Meng, and Hung-yi Lee. Defense against adversarial attacks on spoofing countermeasures of asv. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6564–6568. IEEE, 2020.

[42] Haolin Wu, Jing Chen, Ruiying Du, Cong Wu, Kun He, Xingcan Shang, Hao Ren, and Guowen Xu. Clad: Robust audio deepfake detection against manipulation attacks with contrastive learning. *arXiv preprint arXiv:2404.15854*, 2024.

[43] Yuankun Xie, Yi Lu, Ruibo Fu, Zhengqi Wen, Zhiyong Wang, Jianhua Tao, Xin Qi, Xiaopeng Wang, Yukun Liu, Haonan Cheng, et al. The codecfake dataset and countermeasures for the universally detection of deepfake audio. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

[44] Yuekai Zhang, Ziyan Jiang, Jesús Villalba, and Najim Dehak. Black-box attacks on spoofing countermeasures using transferability of adversarial examples. In *Interspeech*, pages 4238–4242, 2020.

[45] Pedram Abdzadeh Ziabary and Hadi Veisi. A countermeasure based on cqt spectrogram for deepfake speech detection. In *2021 7th International Conference on Signal Processing and Intelligent Systems (ICSPIS)*, pages 1–5. IEEE, 2021.