

THE IMPACT OF AUDIO WATERMARKING ON AUDIO ANTI-SPOOFING COUNTERMEASURES

Zhenshan Zhang¹, Xueping Zhang¹, Yechen Wang², Liwei Jin², Ming Li¹

¹Suzhou Municipal Key Laboratory of Multimodal Intelligent Systems,
Digital Innovation Research Center, Duke Kunshan University, Kunshan, China

²OfSpectrum, Inc., Los Angeles, USA

ABSTRACT

This paper presents the first study on the impact of audio watermarking on spoofing countermeasures. While anti-spoofing systems are essential for securing speech-based applications, the influence of widely used audio watermarking—originally designed for copyright protection—remains largely unexplored. We construct watermark-augmented training and evaluation datasets, named the *Watermark-Spoofing dataset*, by applying diverse handcrafted and neural watermarking methods to existing anti-spoofing datasets. Experiments show that watermarking consistently degrades anti-spoofing performance, with higher watermark density correlating with higher Equal Error Rates (EERs). To mitigate this, we propose the *Knowledge-Preserving Watermark Learning (KPWL)* framework, enabling models to adapt to watermark-induced shifts while preserving their original-domain spoofing detection capability. These findings reveal audio watermarking as a previously overlooked domain shift and establish the first benchmark for developing watermark-resilient anti-spoofing systems. All related protocols are publicly available at https://github.com/Alphawarheads/Watermark_Spoofing.git.

Index Terms— Audio watermarking, Audio Antispoofing, Audio Deepfake Detection, Domain adaptation

1. INTRODUCTION

In recent years, advances in speech synthesis and voice conversion have enabled highly realistic artificial speech, raising serious concerns about security in voice-based applications. Automatic speaker verification (ASV) and its anti-spoofing countermeasures have thus become essential for safeguarding applications such as banking, forensics, and access control [1]. Meanwhile, audio watermarking is widely used for copyright protection and content authentication [2].

Most watermarking research has focused on imperceptibility and robustness to common signal processing operations [3–5], while anti-spoofing has advanced with self-

supervised front ends like XLS-R [6] and WavLM [7], combined with strong back-end models such as AASIST [8], SLS [9] and Nes2Net [10]. However, despite growing research on adversarial perturbations, the impact of audio watermarking—introducing structured yet persistent perturbations—on anti-spoofing systems remains unexplored.

Although audio watermarks are designed to be imperceptible to human listeners, they inevitably modify the underlying speech signal distribution [2, 11], especially in real-world scenarios where multiple watermarking schemes may coexist. Such modifications can introduce complex and previously unstudied domain shifts between original and watermarked audio, potentially leading spoofing detectors to misclassify genuine or synthetic speech. This issue is critical for cross-dataset generalization, where anti-spoofing models must handle unseen spoofing techniques [12], and unexamined interactions with watermarking could compromise their security benefits.

Our main contributions are as follows:

- **Watermark-Spoofing Dataset:** We create dedicated training and evaluation datasets for anti-spoofing models under watermark-induced domain shifts.
- **Extensive evaluation:** A diverse set of handcrafted and DNN-based watermarking methods [5, 13–19] is systematically applied to benchmark datasets [12, 20, 21] to assess their impact on state-of-the-art spoofing detection models.
- **Adaptation framework:** We propose the *Knowledge-Preserving Watermark Learning (KPWL)*, a framework that mitigates watermark-induced performance degradation while preserving original-domain detection accuracy.

2. WATERMARKING IMPACT ANALYSIS

In this section, the impact of audio watermarking on spoofing detection is evaluated. A set of six handcrafted watermarks [11, 15, 17, 19, 22, 23] and three DNN (Deep Neural Network)-based methods [4, 5, 13] is applied to bench-

Corresponding Author: Ming Li, ming.li369@dukekunshan.edu.cn

mark datasets and assessed using three state-of-the-art anti-spoofing models [9, 10, 24]. The Equal Error Rate (EER) of each model is measured as the proportion of watermarked samples in the evaluation data increases. Each watermarked evaluation set maintains a strict 1:1 ratio between DNN-based and handcrafted watermark types, and all watermarking methods within each group are represented in equal proportion. Furthermore, each type of watermark is applied to both bonafide and spoof samples proportionally to their natural distribution within the corresponding dataset. This setup is designed to simulate the cases that both bonafide and spoofing samples could be protected by different kinds of watermarking algorithms. This balanced design ensures that any observed performance differences can be attributed to the presence of watermarking, reducing the effect of class imbalance or unequal representation of specific watermarking methods.

2.1. Watermark-Spoofing Dataset

A mixed watermark dataset comprising six handcrafted methods [11, 15, 17, 19, 22, 23] and three DNN-based methods [4, 5, 13] is constructed to approximate real-world conditions. This collection is applied to three benchmark anti-spoofing datasets—ASVspoof 2021 LA, ASVspoof 2021 DF (LA21, DF21) [20], and the In-the-Wild dataset (ITW) [21]. For each dataset, three watermarked variants are created, with 75%, 50%, and 25% of the samples watermarked while the remaining samples are kept clean, following the distribution described above. Together, these corpora constitute the *Watermark-Spoofing Seen Evaluation dataset*. Similarly, 50% of the ASVspoof 2019 LA (LA19) training data [12] is watermarked using the same configuration, forming the *Watermark-Spoofing Training dataset*.

2.2. Impact of Watermark Data on Anti-Spoofing

Anti-Spoofing on Single Watermark Data: To quantify the impact of individual watermarking methods, four representative DNN-based watermarks—RobustDNN [3], WavMark [4], Timbre [5], and AudioSeal [13]—are applied to the In-The-Wild evaluation dataset (ITW) [21]. Equal Error Rates (EERs) are measured under varying watermark ratios using an XLSR+SLS [9] baseline detector trained on ASVspoof 2019 (LA19) [12].

Table 1 shows that single watermarking indeed degrades spoofing detection performance, with EERs increasing as the watermark proportion grows. Moreover, the observed degradation tends to diminish with newer watermarking methods, suggesting that recent designs introduce less disruptive perturbations.

Anti-Spoofing on Mixed Watermark Data: The *Watermark-Spoofing Seen Evaluation dataset* is then used to assess three state-of-the-art models trained on the LA19 train-

Table 1. EER (%) on ITW dataset under different watermark ratios. Δ shows relative degradation compared to the clean (0%) setting.

Method	75%	50%	25%	0%	75% Δ (%)
AudioSeal (2024) [13]	7.46	7.40	7.35	7.32	1.91%
Timbre (2023) [5]	8.18	7.93	7.53	7.32	11.75%
WavMark (2023) [4]	9.90	9.06	8.23	7.32	35.25%
DNN (2022) [3]	9.06	8.65	8.06	7.32	23.77%

Table 2. EER (%) of three baseline models trained on LA19 and evaluated on *Watermark-Spoofing Seen Evaluation dataset* sets at different ratios.

Model	Test Data	75%	50%	25%	0%	75% Δ (%)
XLSR +AASIST (2022)	LA21	0.88	0.83	0.79	0.73	20.55
	ITW	11.28	10.65	10.00	9.42	19.75
	DF21	6.16	6.08	5.99	5.86	5.12
XLSR +SLS (2024)	LA21	3.68	3.52	3.35	3.02	21.85
	ITW	8.46	7.83	7.57	7.32	15.57
	DF21	2.23	2.13	2.17	2.01	10.94
XLSR +Nes2Net-X (2025)	LA21	2.25	2.20	2.08	2.00	12.50
	ITW	6.40	6.07	5.84	5.50	16.36
	DF21	1.85	1.84	1.82	1.76	5.11

ing set: XLSR-AASIST [24], XLSR-SLS [9], and XLSR-Nes2Net-X [10]. As shown in Table 2, results across all models and datasets reveal a clear positive correlation: higher proportions of watermarked samples consistently lead to increased EERs, demonstrating that watermarking degrades spoofing detection performance.

Furthermore, as shown in Table 3, a multi-dataset [12, 25–27] trained XLSR+SLS model, achieving state-of-the-art results on both experimental (LA21 and DF21) and real-world (ITW) deepfake datasets, also suffers a notable performance drop when tested on watermarked audio. Even the advanced model struggles under watermarked conditions, suggesting that audio watermarking introduces a previously unstudied form of domain shift that current anti-spoofing systems are not designed to adapt to or be robust against.

Limitations of Direct Training for Watermarked Anti-Spoofing: To further examine the domain shift induced by watermarking, the *Watermark-Spoofing Training dataset* was used to train the three selected models—XLSR-AASIST, XLSR-SLS, and XLSR-Nes2Net-X.

As displayed in Table 4, the results were complex: models such as XLSR+SLS and XLSR+Nes2Net-X appeared to extract more stable features and partially suppress watermark-induced shifts as the watermark density increased, albeit at the cost of degraded performance on the original clean evaluation sets.

In contrast, XLSR+AASIST showed more dramatic changes. On the In-The-Wild dataset, it exhibited behavior

Table 3. EER (%) of XLSR+SLS trained on TIMIT [26], FoR [25], ODSS [27] and ASV19 [12] and evaluated under 3 datasets with *Watermark-Spoofing Seen Evaluation dataset*. Δ denotes the relative increase from 0%.

Dataset	75%	50%	25%	0%	75% Δ (%)
LA21	2.82	2.67	2.55	2.46	14.63%
ITW	4.06	3.77	3.49	3.17	28.08%
DF21	1.18	1.10	1.04	1.02	15.69%

Table 4. EER (%) of three models trained with *Watermark-Spoofing Training dataset*, evaluated *Watermark-Spoofing Seen Evaluation dataset*.

Model	Dataset	75%	50%	25%	0%
XLSR +AASIST (2022)	LA21	4.41	4.18	3.96	3.84
	ITW	10.08	9.56	9.16	8.73
	DF21	3.11	2.98	2.85	2.66
XLSR +SLS (2024)	LA21	3.28	3.25	3.23	3.17
	ITW	9.03	8.75	8.57	8.21
	DF21	2.13	2.07	2.00	1.87
XLSR +Nes2Net-X (2025)	LA21	2.89	2.83	2.79	2.75
	ITW	7.58	7.23	6.90	6.58
	DF21	1.92	1.90	1.88	1.87

resembling data augmentation, achieving slight performance gains, while on the ASVspooft2021 LA dataset its performance dropped sharply from a state-of-the-art 0.78% EER to an average level of 3.84%. This degradation may be linked to a reduction of overfitting on LA21 caused by the added watermark noise.

These observations indicate that introducing watermarked audio into training does not provide a consistent benefit and may distort model learning in unexpected ways.

3. METHODOLOGY

As shown in the previous sections, audio watermarking introduces a notable domain shift that degrades the performance of anti-spoofing models. While directly incorporating watermarked data into training can sometimes suppress watermark-induced changes, the results are highly unstable and often compromise the model’s performance on clean, unwatermarked data.

To reduce the negative impact of watermarks while preserving the model’s spoofing detection ability on original data, we propose *Knowledge-Preserving Watermark Learning (KPWL)* as shown in Fig. 1.

Original Baseline Pretraining: We first train the anti-spoofing model on the original LA19 data using a standard supervised objective, establishing a strong original-domain baseline, in this case, the XLSR+SLS model.

Knowledge-Preserving Watermark Learning: Next,

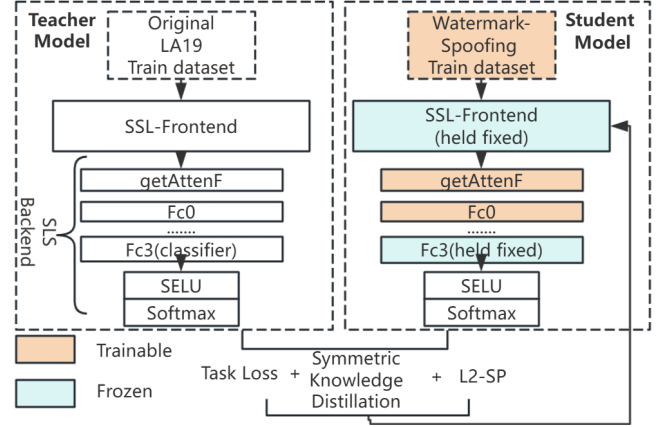


Fig. 1. The Knowledge-Preserving Watermark Learning framework: the model is first trained normally as [9], then adapted with the SSL front end and classifier frozen while only intermediate layers are updated

the *Watermark-Spoofing Training dataset* is used for model adaptation. In this phase, the SSL front end (e.g., XLSR) and the final classification layer (e.g., Fc3 in SLS) are kept frozen (non-trainable), while only the intermediate backend layers are updated. Training is conducted with a slightly higher learning rate than in the previous phase and for 2 epochs, enabling rapid adaptation with minimal drift. To stabilize this process, a teacher–student framework is employed, incorporating symmetric knowledge distillation and parameter anchoring (L2-SP), as shown in Eq. 1.

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{task}}}_{\text{supervised}} + \beta \underbrace{\mathcal{L}_{\text{KD}}}_{\text{teacher-student}} + \mu \underbrace{\mathcal{L}_{\text{L2-SP}}}_{\text{parameter anchor}}, \quad (1)$$

$$\mathcal{L}_{\text{KD}} = D_{\text{KL}}(p_T \| p_S) + D_{\text{KL}}(p_S \| p_T),$$

$$\mathcal{L}_{\text{L2-SP}} = \sum_{i \in \mathcal{I}_{\text{trainable}}} \|w_i - w_i^{(0)}\|_2^2,$$

where $\mathcal{L}_{\text{task}}$ is a weighted cross-entropy loss to optimize classification performance. \mathcal{L}_{KD} is a symmetric knowledge distillation loss that constrains the adapted model’s predictions to remain close to those of the frozen teacher snapshot. In addition, L2-SP regularization $\mathcal{L}_{\text{L2-SP}}$ penalizes deviation of trainable parameters from their initialization at the start of Knowledge-Preserving Watermark Learning phase, thereby reducing parameter drift during adaptation.

Freezing the SSL front end preserves low-level features against watermark-induced shifts, while fixing the classifier maintains the decision boundary. Symmetric Knowledge Distillation constrains the *outputs* to remain close to the pre-adaptation model, while L2-SP constrains the *trainable parameters* not to drift far from the Knowledge-Preserving Watermark Learning phase starting point. Together, these choices allow the model to maintain its original-domain performance while reliably adapting to watermarked inputs.

4. EXPERIMENTAL ANALYSIS

4.1. Experimental Settings

Preprocessing: All audio is resampled to 16 kHz and normalized to 64,600 samples by tiling or truncation. RawBoost [28] is applied with colored noise at random SNRs, and the resulting waveforms are converted to single-channel format.

Training: Original Baseline Pretraining phase trains on the LA19 split using Adam with a learning rate of 1×10^{-7} , and weight decay of 10^{-4} , optimized with class-weighted cross-entropy for up to 50 epochs with early stopping. The KPWL adaptation phase continues from Original Baseline Pretraining phase, trained with *Watermark-Spoofing Training dataset*, the watermarked version of LA19, using Adam with a learning rate of 5×10^{-7} , and weight decay of 10^{-4} . The SSL frontend and classifier are frozen while only intermediate layers are updated, trained for 2 epochs using class-weighted NLL with symmetric knowledge distillation ($\beta = 0.3$) and L2-SP regularization ($\mu = 10^{-4}$).

Evaluation: To assess the effectiveness of the KPWL, three variations of XLSR+SLS were compared: (i) Model trained solely on original LA19 data (hereafter referred to as the baseline model), (ii) Model trained on *Watermark-Spoofing Training dataset* (hereafter referred to as the watermarked model), and (iii) Model trained using the KPWL method (hereafter referred to as the KPWL model). These models were evaluated on the LA21, DF21, and *In-the-Wild* datasets, each tested in their original form and in watermarked versions—collectively referred to as the *Watermark-Spoofing Seen Evaluation dataset*—as well as on a more challenging *Watermark-Spoofing Unseen Evaluation dataset*, which was generated using the same procedure as the seen set but with a distinct set of watermarking methods [3, 11, 14, 16, 18, 29].

4.2. Experimental Results

As shown in Table 5, the proposed KPWL model consistently outperforms the baseline on the *Watermark-Spoofing Seen Evaluation dataset*, reducing the EER on ITW with 75% watermarked data from 8.46% to 7.92%. It also maintains comparable performance on clean data (3.06% vs. 3.02% on LA21). In contrast, the watermarked model gains robustness against domain shift as watermark density increases, but sacrifices clean-domain accuracy. These results show that KPWL suppresses watermark-induced domain shifts while preserving spoofing detection capability, indicating that existing models can adapt to watermarked data when guided by an appropriate training strategy.

As shown in Table 6, performance changes were far more drastic on the *Watermark-Spoofing Unseen Evaluation dataset*. The baseline model outperformed both the watermarked and KPWL variants (9.94% vs. 10.21% and 11.22% on 75%-watermarked data from the *Watermark-Spoofing Seen Evaluation dataset* LA21), indicating that the latter two may

Table 5. Comparison of three XLSR+SLS variants on different test datasets under *Watermark-Spoofing Seen Evaluation dataset*. Δ denotes relative increase from 0%.

Dataset	Model	75%	50%	25%	0%
LA21	Baseline	3.68	3.52	3.35	3.02
	Watermarked	3.28	3.25	3.23	3.17
	KPWL	3.21	3.18	3.12	3.06
ITW	Baseline	8.46	7.83	7.57	7.32
	Watermarked	9.03	8.75	8.57	8.21
	KPWL	7.92	7.74	7.60	7.37
DF21	Baseline	2.23	2.13	2.17	2.01
	Watermarked	2.13	2.07	2.00	1.87
	KPWL	2.04	1.95	1.92	1.74

Table 6. EER(%) of different variants of XLSR+SLS under *Watermark-Spoofing Unseen Evaluation dataset* across different test datasets.

Dataset	Model	75%	50%	25%	0%
LA21	Baseline	9.94	8.01	5.72	3.07
	Watermarked	10.21	8.00	5.64	3.17
	KPWL	11.22	8.47	5.66	3.04
ITW	Baseline	13.41	11.43	9.36	7.32
	Watermarked	15.66	13.15	10.68	8.62
	KPWL	14.78	12.27	9.71	7.37
DF21	Baseline	8.42	6.34	4.31	2.01
	Watermarked	8.60	6.40	4.20	1.87
	KPWL	10.05	7.24	4.50	1.75

have over-adapted to the specific watermarking methods used to generate the *Watermark-Spoofing Training dataset*. This suggests that current anti-spoofing models lack robustness to diverse distortions, with watermarking exposing a critical gap that warrants further investigation.

5. CONCLUSION

This work presents the first study on how audio watermarking impacts anti-spoofing detection systems. Experiments across diverse datasets and models show that watermarking can significantly degrade detector performance during both training and evaluation. To support future research, we constructed watermark-augmented training and evaluation datasets. To mitigate mentioned issues, we propose KPWL, a teacher-student adaptation framework that preserves original-domain decision boundaries while adapting to watermark-induced shifts. These findings highlight audio watermarking as an overlooked source of domain shift and underscore the need for anti-spoofing models robust to real-world watermark contamination, while handling unseen watermark scenarios remains an open challenge.

6. ACKNOWLEDGEMENTS

This research is funded by the DKU Foundation Project “Emerging AI Technologies for Natural Language Processing”. We also thank the Advanced Computing East China Sub-Center for providing computational resources.

7. REFERENCES

- [1] Andre Kassis, Urs Hengartner, “Practical attacks on voice spoofing countermeasures,” *CoRR*, vol. abs/2107.14642, 2021.
- [2] Guang Hua, Jiwu Huang, Yun Q. Shi, et al., “Twenty years of digital audio watermarking—a comprehensive review,” *Signal Processing*, vol. 128, pp. 222–242, 2016.
- [3] K. Pavlović et al., “Robust speech watermarking by a jointly trained embedder and detector using a dnn,” *Digital Signal Processing*, vol. 122, pp. 103381, 2022.
- [4] Guangyu Chen, Yu Wu, Shujie Liu, et al., “Wavmark: Watermarking for audio generation,” 2023.
- [5] Chang Liu, Jie Zhang, Tianwei Zhang, et al., “Detecting voice cloning attacks via timbre watermarking,” in *Network and Distributed System Security Symposium*, 2024.
- [6] Zhiqiang Lv, Shanshan Zhang, Kai Tang, et al., “Fake audio detection based on unsupervised pretraining models,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 9231–9235.
- [7] S. Chen, C. Wang, Z. Chen, et al., “WavLM: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [8] J. w. Jung, H.-S. Heo, H. Tak, et al., “AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 6367–6371.
- [9] Qishan Zhang, Tao Hu, “Audio deepfake detection with self-supervised xls-r and sls classifier,” in *Proc. ACM International Conference on Multimedia*, 2024, p. 6765–6773.
- [10] Tianchi Liu, Duc-Tuan Truong, Rohan Kumar Das, et al., “Nes2net: A lightweight nested architecture for foundation model driven speech anti-spoofing,” *arXiv preprint arXiv:2504.05657*, 2025.
- [11] A. A. Alsabghany, F. Ridzuan et al., “Digital audio steganography: Systematic review, classification, and analysis of the current state of the art,” *Computer Science Review*, vol. 38, pp. 100316, 2020.
- [12] X. Wang, J. Yamagishi, M. Todisco, et al., “ASVspoof 2019: A large-scale public database of synthesized, converted and re-played speech,” *Computer Speech & Language*, vol. 64, pp. 101114, 2020.
- [13] Robin San Roman, Pierre Fernandez, Hady Elsahar, et al., “Proactive detection of voice cloning with localized watermarking,” in *Proc. ICML*, 2024, vol. 235, pp. 1–17.
- [14] Mayank Kumar Singh, Naoya Takahashi, Weihsiang Liao, et al., “Silentcipher: Deep audio watermarking,” in *Proc. Interspeech*, 2024, pp. 2235–2239.
- [15] J. Zhao, T. Zong et al., “Desynchronization attacks resilient watermarking method based on frequency singular value coefficient modification,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2282–2295, 2021.
- [16] Rizky M. Nugraha, “Implementation of direct sequence spread spectrum steganography on audio data,” in *International Conference on Electrical Engineering and Informatics*, 2011.
- [17] Pranab Kumar Dhar, Tetsuya Shimamura, “Blind svd-based audio watermarking using entropy and log-polar transformation,” *Journal of Information Security and Applications*, vol. 20, pp. 74–83, 2015.
- [18] F. A. Adhiyaksa, T. Ahmad, A. M. Shiddiqi, et al., “Reversible audio steganography using least prime factor and audio interpolation,” in *International Seminar on Machine Learning, Optimization, and Data Science*, 2022, pp. 97–102.
- [19] S. Mushtaq, S. A. Parah, “Blind and robust watermarking framework for audio signals,” in *International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)*, 2024, pp. 1–5.
- [20] J. Yamagishi, X. Wang, M. Todisco, et al., “ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection,” in *Automatic Speaker Verification and Spoofing Countermeasures (ASVspoof) Challenge*, 2021, pp. 47–54.
- [21] N. M. Müller, P. Czempin, F. Dieckmann, et al., “Does audio deepfake detection generalize?,” in *Proc. INTERSPEECH*, 2022, pp. 2783–2787.
- [22] Guang Yang, “An improved phase coding audio steganography algorithm,” in *Journal of The Colloquium for Information Systems Security Education*, 2025, vol. 12, pp. 1–4.
- [23] Iynkaran Natgunanathan, Yong Xiang, Guang Hua, et al., “Patchwork-based multilayer audio watermarking,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2176–2187, 2017.
- [24] Hemlata Tak, Massimiliano Todisco, Xin Wang, et al., “Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation,” in *The Speaker and Language Recognition Workshop*, 2022.
- [25] R. Reimao, V. Tzerpos, “FoR: A dataset for synthetic speech detection,” in *Proc. International Conference on Speech Technology and Human-Computer Dialogue*, 2019.
- [26] D. Salvi, B. Hosler, P. Bestagini, et al., “TIMIT-TTS: A text-to-speech dataset for multimodal synthetic media detection,” *IEEE Access*, vol. 11, 2023.
- [27] A. Yaroshchuk, C. Papastergiopoulos, L. Cuccovillo, et al., “An Open Dataset of Synthetic Speech,” in *IEEE International Workshop on Information Forensics and Security*, 2023.
- [28] Hemlata Tak, Madhu Kamble, Jose Patino, et al., “RawBoost: A raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 6382–6386.
- [29] Slami Saadi, Ali Benziane, “Novel secured scheme for blind audio/speech norm-space watermarking by arnold algorithm,” *Signal Processing*, vol. 154, pp. 74–86, 2019.