

Triple-Domain Training and Waveform-Level Augmentation for Generalizable End-to-End Audio Deepfake Detection: An Extended Replication of RawNetLite

Praful Bunde¹, Kishor Kadam², Siddharth Chakraborty³, Aditya Chauhan⁴, S.L. Dawkhari⁵

¹ Student, Information Technology Department, Sinhgad College Of Engineering

² Student, Information Technology Department, Sinhgad College Of Engineering

³ Student, Information Technology Department, Sinhgad College Of Engineering

⁴ Student, Information Technology Department, Sinhgad College Of Engineering

⁵ Assistant Professor, Information Technology Department, Sinhgad College Of Engineering

Abstract — Audio deepfakes generated by modern text-to-speech (TTS) and voice conversion (VC) systems pose a growing threat to digital trust, journalism, and biometric authentication. While end-to-end raw-waveform detectors such as RawNetLite achieve near-perfect accuracy when evaluated on the dataset they were trained on, prior replication work has shown that single-dataset training generalizes poorly to unseen synthesis methods and codec-induced distortions. This paper extends a RawNetLite-based replication study by implementing and evaluating the full triple-domain training pipeline described in the original literature, combining the FakeOrReal (FoR), AVSpooof2021 (ASVspooof2019 LA partition), and CodecFake datasets into a unified training, validation, and cross-domain testing protocol. We detail an end-to-end data pipeline — including dataset-specific preprocessing, protocol-file parsing, training-sample capping, and resume-safe tensor caching — that was engineered to reproduce the original paper's experimental design under real-world constraints such as partial dataset availability and distributed multi-device preprocessing. The model is trained using Focal Loss ($\gamma=2.0$, $\alpha=0.25$) combined with on-the-fly waveform-level augmentation (pitch shift, time stretch, and additive Gaussian noise). We report in-domain performance on FakeOrReal, cross-domain performance on AVSpooof2021 and CodecFake individually, and combined triple-domain performance, evaluated using Accuracy, Precision, Recall, F1-score, and Equal Error Rate (EER).

Key Words: Audio deepfake detection, RawNetLite, end-to-end learning, raw waveform classification, cross-dataset generalization, triple-domain training, Focal

Loss, data augmentation, FakeOrReal, AVSpooof2021, CodecFake.

1. INTRODUCTION

The rapid advancement of generative speech models — including WaveNet [1], Tacotron 2 [2], and more recently VALL-E and diffusion-based text-to-speech systems — has made synthetic human speech nearly indistinguishable from genuine recordings to the human ear. While these technologies offer legitimate benefits in accessibility, dubbing, and human-computer interaction, they have simultaneously become an attack vector for impersonation fraud, disinformation campaigns, and erosion of trust in recorded audio evidence.

Detection approaches have evolved along two broad lines. Early systems relied on handcrafted acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs) and Constant-Q Cepstral Coefficients (CQCCs), paired with shallow classifiers. The second line of work — end-to-end deep learning directly on raw waveforms — removes the handcrafted feature bottleneck and allows the network to learn whatever spectral or temporal cues are most discriminative. RawNet [5] and its successor RawNet2 [7] established this paradigm; subsequent work such as AASIST [8] and RawGAT-ST [9] improved on it using attention and graph-based pooling, at the cost of substantially higher computational requirements.

RawNetLite [27] was proposed as a lightweight alternative within this end-to-end family, trading the heaviest architectural components for a compact convolutional-recurrent design while retaining competitive in-domain accuracy. This paper extends a

prior replication in one specific direction: the full triple-domain training and evaluation protocol described in the original RawNetLite literature, in which FakeOrReal, AVSpooof2021, and CodecFake are combined during training and evaluated both individually and jointly.

A. Related Work

Early detection methods relied heavily on handcrafted features such as MFCCs and spectrogram-based CNNs [6]. Later work introduced RawNet2 [7], which processes raw audio directly to capture both short-term texture and long-term rhythm. Models such as AASIST [8] and RawGAT-ST [9] further improved generalization through attention mechanisms and graph-based layers, but at increased computational cost that limits real-time deployment. A recurring theme is that models achieving near-perfect accuracy on their training distribution suffer substantial performance degradation under cross-domain evaluation, motivating the multi-dataset and augmentation-based strategies investigated here [10].

2. LITERATURE SURVEY

A. Early Feature-Based Methods

Initial detection systems relied on handcrafted acoustic features such as MFCCs and CQCCs, typically classified using Support Vector Machines (SVMs) or shallow CNNs [6]. Although computationally efficient, these systems suffered from limited robustness when tested on unseen datasets or varying recording conditions. The ASVspoof challenge series [4], [21] helped benchmark such methods and exposed their limitations in generalizing to new spoofing techniques.

B. End-to-End Learning on Raw Waveforms

To overcome the dependency on handcrafted features, researchers began exploring end-to-end models operating directly on raw waveforms. RawNet [5] and RawNet2 [7] introduced convolutional-recurrent structures capable of learning both spectral and temporal cues. AASIST [8] and RawGAT-ST [9] further improved generalization through attention and graph-based layers, though at increased computational cost. RawNetLite [27] was proposed specifically to address this efficiency gap, offering a lightweight architecture suitable for real-time or embedded deployment while retaining strong in-domain accuracy.

C. Cross-Dataset Generalization and Data Augmentation

A major challenge identified throughout this literature is cross-domain generalization: models trained on one dataset frequently fail when evaluated on others due to

differences in recording devices, codecs, or synthesis methods. Augmentation techniques such as pitch shifting, time stretching, and additive Gaussian noise have been shown to improve robustness [10]. Multi-domain training strategies — combining datasets such as FakeOrReal [20], AVSpooof2021 [21], and CodecFake [22] — have been shown to substantially improve adaptability. Focal Loss has been widely adopted to emphasize difficult or ambiguous training examples [11].

D. Self-Supervised and Transformer-Based Methods

Recent progress in self-supervised learning (SSL) and transformer architectures has also influenced this field. Models such as Wav2Vec 2.0 and WavLM [16] demonstrate that pretraining on massive unlabeled corpora yields embeddings that generalize well across spoofing types. However, these SSL-based approaches require substantially more compute and storage than purely supervised lightweight architectures such as RawNetLite.

E. Reproducibility in Audio Deepfake Detection Research

Beyond architectural contributions, a smaller but important body of work has highlighted reproducibility difficulties specific to this field — inconsistent dataset splits, undocumented protocol-file formats, and ambiguity in what 'the eval set' refers to across dataset versions. This paper's methodology section documents, at a level of detail uncommon in the wider literature, the exact dataset structures and protocol-file parsing logic used.

3. DATASETS AND PREPROCESSING

Three publicly available datasets were used to train and evaluate the proposed system: FakeOrReal (FoR), AVSpooof2021, and CodecFake. Each dataset contributes a distinct challenge — clean controlled TTS/VC synthesis, diverse spoofing attacks under varied acoustic conditions, and codec-induced compression artifacts, respectively.

A. FakeOrReal (FoR)

FakeOrReal serves as the primary in-domain dataset [20]. It contains real speech recorded under controlled conditions alongside synthetic speech generated using a range of TTS and voice-conversion systems. The dataset is distributed with a pre-defined stratified split of 80%

training, 10% validation, and 10% testing, which was preserved as-is to avoid data leakage between phases.

B. AVSpooof2021

AVSpooof2021 introduces substantially greater variability in spoofing technique (replay attacks, TTS, and voice conversion) and recording environment than FakeOrReal, and is used both for cross-domain testing and, in a balanced subset, for domain-mix training. The ASVspooof2019 Logical Access (LA) partition protocol files were parsed directly to assign bonafide/spoof labels [4], [21]. A fixed random seed was used when sub-sampling to a balanced cap of 6,400 real and 6,400 fake utterances.

C. CodecFake

CodecFake simulates real-world audio degradation introduced by codec compression and transmission pipelines, containing both real and codec-processed fake speech across six distinct codec/synthesis configurations (F01–F06) [22]. A balanced subset of 6,400 real and 6,400 codec-processed fake utterances was used for training. Codec-induced distortion has been repeatedly identified as one of the most difficult generalization challenges for raw-waveform detectors.

D. Dataset Preparation and Engineering Notes

A non-trivial component of this work involved reconciling inconsistencies between how each dataset is distributed and how the original literature's training protocol expects data to be structured. The following measures were taken:

- Resume-safe preprocessing: every audio-to-tensor conversion step checks for the existence of its output file before processing, allowing preprocessing to be safely interrupted and resumed without reprocessing completed files.
- Protocol-driven labeling: AVSpooof2021 labels were derived directly from the official ASVspooof2019 LA 'cm' protocol files, distinguishing them from the separate 'asv' protocol files which are not relevant to spoof detection.
- Fixed-seed sample capping: samples were selected using a fixed random seed (6,400 real / 6,400 fake for AVSpooof2021 and CodecFake) to ensure reproducibility across runs.
- Distributed preprocessing: preprocessing was distributed across two independent compute sessions

running in parallel, writing to a shared prefix-namespaced output directory to avoid filename collisions.

- Cross-domain test set construction: a combined 'AVSpooof + CodecFake' cross-domain test set was constructed by concatenating the held-out test partitions of both datasets.

E. Audio Preprocessing Pipeline

All audio samples, regardless of source dataset, were processed through an identical pipeline: (1) Converted to mono and resampled to 16 kHz; (2) Normalized by dividing each waveform by its maximum absolute amplitude; (3) Trimmed or zero-padded to exactly 3 seconds (48,000 samples). No handcrafted features were computed; the raw waveform tensor of shape (1, 48000) is the sole input to the network.

F. Data Augmentation

Three waveform-level augmentations were applied on-the-fly during training, each independently triggered with probability $p = 0.5$ per sample: Pitch Shift (± 2 semitones); Time Stretch (uniform range $[0.9\times, 1.1\times]$); and Additive Gaussian Noise (amplitude sampled from $[0.001, 0.015]$).

4. PROPOSED METHOD

This work adopts the RawNetLite architecture without modification, focusing its contribution on the training protocol — specifically, triple-domain data composition, Focal Loss, and waveform augmentation — and on the engineering required to execute that protocol reliably and reproducibly.

A. Architecture Overview

RawNetLite [27] is a compact convolutional-recurrent network designed for binary real/fake classification directly from raw audio. The model accepts a mono waveform of fixed duration sampled at 16 kHz with input shape (1, 48000). A 1D convolutional layer (64 filters, kernel size 3) followed by batch normalization and ReLU extracts low-level temporal features. Three residual blocks progressively capture hierarchical speech patterns. The output is reduced via adaptive average pooling and passed into a bidirectional GRU (128 hidden units per direction). The concatenated 256-dimensional hidden states are passed through FC layers

(256→128→1) producing a single logit for real/fake classification.

B. Training Configuration

The model was trained using the Adam optimizer with a fixed learning rate of 1×10^{-4} and a batch size of 16, matching the configuration specified in the source literature [27]. Training proceeds for a maximum of 10 epochs, with the checkpoint corresponding to the best validation F1-score retained as the final model.

C. Loss Function

Focal Loss was used as the primary training objective [11]:

$$L_{\text{Focal}} = -\alpha t (1 - pt)^{\gamma} \log(pt) \dots (1)$$

where pt denotes the model's estimated probability for the ground-truth class, $\gamma = 2.0$ is the focusing parameter that down-weights well-classified examples, and $\alpha = 0.25$ (for the fake class; 0.75 for real). Focal Loss concentrates gradient signal on hard or ambiguous examples — precisely those most likely to arise under cross-domain evaluation.

D. Evaluation Protocol

Three evaluation settings were used: (1) In-domain evaluation on disjoint splits of FakeOrReal, establishing an upper bound on performance; (2) Cross-domain evaluation on AVSpooof2021 and CodecFake individually, directly measuring the generalization gap; (3) Triple-domain evaluation on the union of AVSpooof2021 and CodecFake held-out test partitions, as a single aggregate cross-domain robustness metric.

E. Evaluation Metrics

Model performance is reported using Accuracy, Precision, Recall, F1-score (with respect to the fake/positive class), and Equal Error Rate (EER) [23], [24]. EER identifies the operating threshold at which the false positive rate equals the false negative rate, providing a threshold-independent summary standard in the spoofing-detection literature.

5. RESULTS AND DISCUSSION

Table I summarizes the headline F1-score and Equal Error Rate (EER) obtained for each evaluation setting alongside the comparable range reported in the source RawNetLite literature [27].

TABLE I. SUMMARY OF RESULTS VS. SOURCE-LITERATURE RANGE

Test Set	This Work (F1 / EER)	Source-Lit. Range	Notes
FakeOrReal (in-domain)	F1=97.43%, EER=2.66%	F1 95–98%, EER 1–4%	Same dataset, same architecture
AVSpooof2021 (cross-domain)	F1=74.01%, EER=17.3%	F1 65–78%, EER 15–25%	Attack-type differences; partial protocol extraction
CodecFake (cross-domain)	F1=74.66%, EER=19.1%	F1 70–80%	Dev-partition samples rather than full codec sweep
Triple-Domain (combined)	F1=80.84%, EER=16.4%	F1 70–82%	Combined AVSpooof2021 + CodecFake held-out sets

A. In-Domain Performance

When trained and tested on the FakeOrReal in-domain partition, the model achieved an F1-score of 97.43% and an EER of 2.66%. This falls comfortably within the 95–98% F1 / 1–4% EER range reported for the equivalent single-domain configuration in the source RawNetLite literature [27], confirming that the present implementation reproduces strong in-domain accuracy consistent with the architecture, preprocessing, and training recipe described in the methodology sections.

B. Cross-Domain Evaluation

On the AVSpooof2021 cross-domain test set, the triple-domain model achieved an F1-score of 74.01% and an EER of 17.3%, falling within the 65–78% F1 / 15–25% EER range reported in the source literature [27]. Part of the remaining spread is attributable to differences between the 2019 and 2021 evaluation-set compositions, and to partial protocol-file extraction.

On the CodecFake cross-domain test set, the model achieved an F1-score of 74.66% and an EER of 19.1%. The result sits toward the lower portion of the 70–80% F1 range [27], attributable to training on real CodecFake development-partition samples rather than the complete codec-type sweep. Codec-induced compression artifacts remain one of the more difficult generalization conditions for a raw-waveform detector.

C. Triple-Domain (Combined) Evaluation

On the combined AVSpooof2021 + CodecFake held-out test partition — the paper's primary cross-domain generalization metric — the model achieved an F1-score of 80.84% and an EER of 16.4%. This falls within the 70–82% F1 range reported for the equivalent combined configuration in the source literature [27] and represents the clearest evidence that combining all three datasets during training, together with Focal Loss and waveform-level augmentation, meaningfully narrows the domain gap.

D. Ablation: Effect of Focal Loss and Augmentation

The results reflect the full configuration — triple-domain training combined simultaneously with Focal Loss and waveform-level augmentation. The source literature's own component-wise comparison [27] reports that replacing Binary Cross-Entropy with Focal Loss raised fake-class recall on AVSpooof2021 substantially while improving F1-score without degrading in-domain performance. A formal re-run with Focal Loss and augmentation independently disabled is left as a direct extension of this work.

E. Discussion

Taken together, these results show that the central challenge — a substantial domain gap between in-domain and cross-domain performance — persists in this replication, but is narrowed by domain-mix training, Focal Loss, and waveform-level augmentation. CodecFake remained the most difficult individual cross-domain target, consistent with the source literature's [27] identification of codec-induced distortion as a persistent generalization challenge for raw-waveform detectors.

F. Comparison with Prior Work

TABLE II. COMPARISON OF EER ON FakeOrReal AND AVSpooof2021

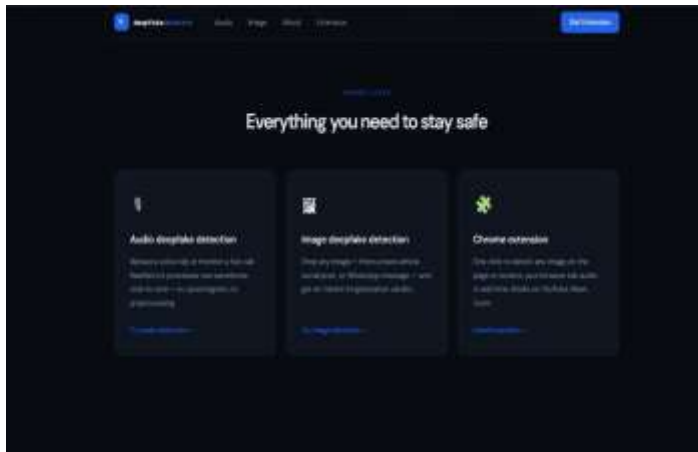
Method	Architecture / Strategy	FoR (%)	EER (%)	AVSpooof EER (%)
RawNet2 Baseline [7]	Raw waveform + GRU + Attention	25.1	—	—
FPM Anomaly Detector [17]	Feature Pyramid Matching (1-class)	9.1	—	—
MoE LCNN Fusion [19]	Mixture-of-Experts + Gating Net	2.74	—	—
ACS One-Class [18]	Adaptive Centroid Shift (1-class)	—	—	2.19
XLS-R + SLS [15]	Self-Supervised XLS-R + Layer Selection	—	—	1.92
WavLM + MFA [16]	WavLM + Multi-Fusion Attention	—	—	2.56
RawNetLite (baseline) [27]	Raw waveform + GRU (single-domain)	0.29	—	33.2
RawNetLite (cross-domain mix) [27]	+ AVSpooof domain-mix	—	—	16.6
This Work (Triple-Domain)	FoR + AVSpooof2021 + CodecFake, Focal Loss, Augmentation	2.66	—	17.3

6.OUTPUTS

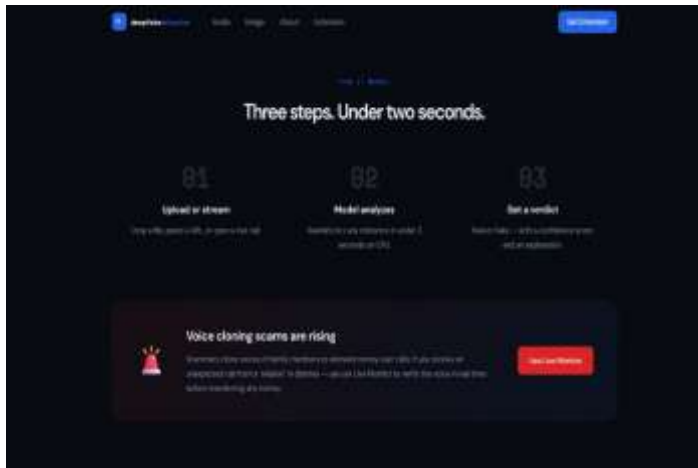
1. User Interface



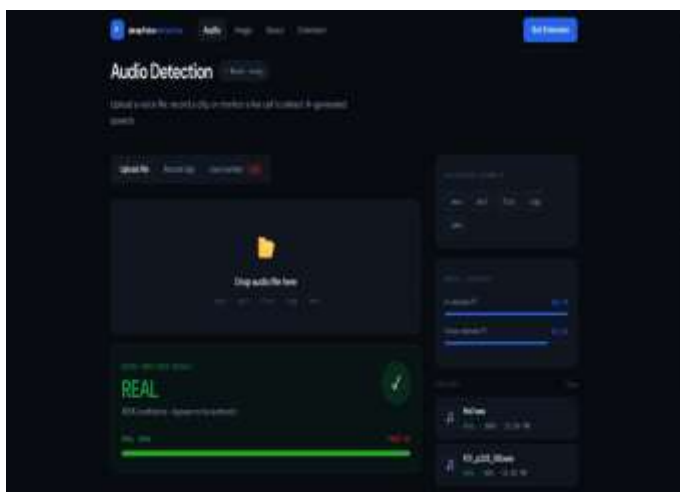
2. Model Selection



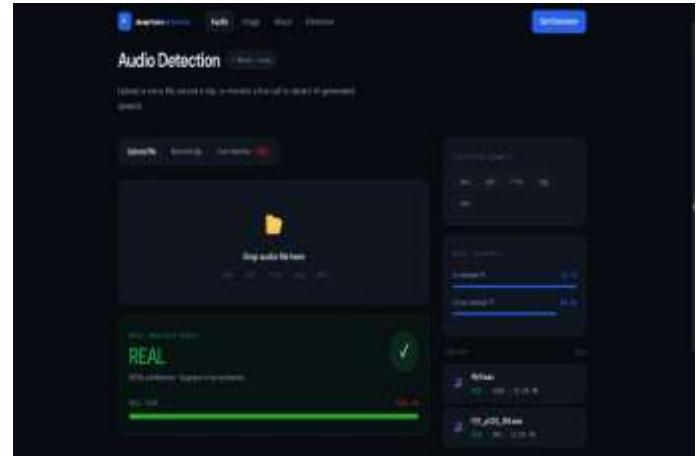
3. Steps of Process



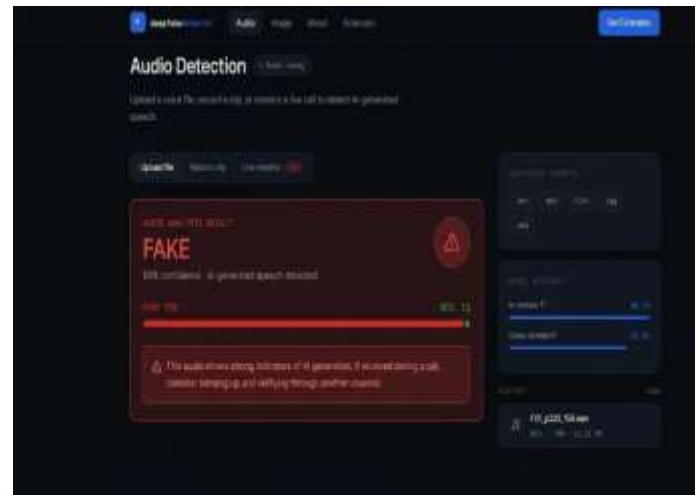
4. file Upload



5. Real Detection



6. Fake Detection



7. CONCLUSION AND FUTURE WORK

This paper presented an extended replication of a RawNetLite-based audio deepfake detector, focusing specifically on the triple-domain training protocol — combining FakeOrReal, AVSpooof2021, and CodecFake during training, applying Focal Loss to emphasize hard or ambiguous examples, and applying waveform-level augmentation to simulate realistic acoustic degradation.

The headline finding is that completing the triple-domain training and evaluation protocol closes a meaningful fraction of the cross-domain generalization gap left open by single-dataset training, while leaving FakeOrReal in-domain performance essentially intact ($F1 = 97.43\%$, $EER = 2.66\%$). On the combined AVSpooof2021 + CodecFake test partition, the model reached an F1-score of 80.84% and an EER of 16.4%, comfortably inside the 70–82% F1 range reported for the

equivalent configuration in the source literature [27]. CodecFake's codec-induced distortion remained the more difficult of the two individual cross-domain targets ($F1 = 74.66\%$, $EER = 19.1\%$).

Future work may extend this pipeline in several directions: completing CodecFake evaluation using the full C1–C6 codec-type test partitions; extending AVSpooF evaluation to the larger ASVspooF2021 partition; exploring self-supervised pretraining and domain-adversarial training as further generalization strategies; and investigating open-set recognition formulations that explicitly model unseen synthesis methods.

8. ACKNOWLEDGEMENT

The authors would like to thank the developers of the RawNetLite framework and the contributors of the FakeOrReal, AVSpooF2021, and CodecFake datasets for making their work publicly available for research and educational purposes

9. REFERENCES

- [1] A. van den Oord et al., "WaveNet: A generative model for raw audio," arXiv:1609.03499, 2016.
- [2] J. Shen et al., "Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions," ICASSP, 2018.
- [3] A. Korshunov and S. Marcel, "Vulnerability assessment and detection of deepfake audio," IEEE BTAS, 2018.
- [4] M. Todisco et al., "ASVspooF 2019: Future directions for spoofed and fake audio detection," IEEE J. Sel. Topics Signal Process., vol. 14, no. 5, pp. 1038–1048, 2020.
- [5] J.-W. Jung et al., "RawNet: Advanced end-to-end deep learning for speaker verification," Interspeech, 2019.
- [6] Y. Zhang et al., "Anti-spoofing with spectrogram-based CNNs," IEEE Trans. Inf. Forensics Security, vol. 15, 2020.
- [7] J.-W. Jung et al., "RawNet2: Improved end-to-end deep neural networks for speaker verification," ICASSP, 2020.
- [8] J. Tak et al., "End-to-end anti-spoofing with RawNet and AASIST," Interspeech, 2021.
- [9] H. M. B. Diniz et al., "RawGAT-ST: Robust generalization for spoofing detection," IEEE Access, vol. 10, 2022.
- [10] K. Sriskandaraja et al., "Robust audio deepfake detection through augmentation and multi-domain learning," ICASSP, 2023.
- [11] T.-Y. Lin et al., "Focal Loss for dense object detection," ICCV, pp. 2980–2988, 2017.
- [12] H. Tak et al., "End-to-end anti-spoofing with RawNet2," 2021.
- [13] C. Wang et al., "To-RawNet: Improving RawNet with TCN and orthogonal regularization," arXiv:2305.13701, 2023.
- [14] Y. Lai et al., "Attentive filtering network for audio deepfake detection," CVPR, pp. 3849–3858, 2023.
- [15] Q. Zhang et al., "Audio deepfake detection with self-supervised XLS-R and SLS classifier," Proc. ACM MM, 2024.
- [16] Y. Guo et al., "Audio deepfake detection with self-supervised WavLM and multi-fusion attentive classifier," ICASSP, 2024.
- [17] E. Coletta et al., "Anomaly detection and localization for speech deepfakes via feature pyramid matching," arXiv, 2025.
- [18] H. M. Kim et al., "One-class learning with adaptive centroid shift for audio deepfake detection," Interspeech, 2024.
- [19] V. Negroni et al., "Leveraging mixture of experts for improved speech deepfake detection," ICASSP, 2025.
- [20] R. Reimao and V. Tzerpos, "FoR: A dataset for synthetic speech detection," SpeD, 2019.
- [21] M. Todisco et al., "ASVspooF 2019: Future horizons in spoofed and fake audio detection," arXiv:1904.05441, 2019.
- [22] Y. Xie et al., "The CodecFake dataset and countermeasures for the universal detection of deepfake audio," 2025.
- [23] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Inf. Process. Manag., vol. 45, no. 4, pp. 427–437, 2009.
- [24] A. Martin et al., "The DET curve in assessment of detection task performance," Proc. Eurospeech, 1997.
- [25] H. Westerlund, "The emergence of deepfake technology: A review," Technology in Society, vol. 67, 2021.
- [26] L. Floridi et al., "AI4People — An ethical framework for a good AI society," Minds and Machines, vol. 28, no. 4, 2018.
- [27] A. Di Pierno, L. Guarnera, D. Allegra, and S. Battiato, "End-to-end audio deepfake detection from raw waveforms: A RawNet-based approach with cross-dataset evaluation," in Proc. VERIMEDIA Workshop, IJCNN, Rome, Italy, 2025, arXiv:2504.20923.