International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 11 | Nov - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

A Lightweight End-to-End Model for Detecting Audio Deepfakes Using Raw Waveforms

Praful Bundele, Kishor Kadam, Siddharth Chakraborty, Aditya Chauhan, S.L.Dawkhar

Department : Information Technology
College: Sinhgad College Of Engineering

Abstract - With the rise of generative artificial intelligence, synthetic voices—commonly known as audio deepfakes—have become increasingly convincing and difficult to distinguish from genuine recordings. These technologies pose major threats to privacy, media integrity, and digital trust. This study replicates and analyzes an existing **RawNetLite**-based approach for end-to-end audio deepfake detection using raw waveforms. The model eliminates manual feature extraction, directly processing 1D audio signals through convolutional and recurrent neural layers. Cross-dataset experiments were conducted using the **FakeOrReal**, AVSpoof2021, and **CodecFake** datasets to assess the model's robustness. The results show strong performance on in-domain data and notable generalization improvements when combining diverse datasets and applying audio augmentation. This paper provides both a technical evaluation and a broader reflection on the ethical and social implications of audio **deepfake** technologies

Key Words: Deepfake detection, audio forensics, RawNetLite, end-to-end learning, cross-dataset evaluation, focal loss, synthetic speech, deep learning.

1.INTRODUCTION

Advancements in deep generative models such as **WaveNet**, **Tacotron** 2, and VALL-E have enabled machines to synthesize human speech with near-perfect naturalness [1], [2]. While these models facilitate accessibility and entertainment, they also enable malicious uses such as identity theft, misinformation, and fraud. Audio deepfakes—synthetic voices generated to imitate real individuals—pose significant challenges for authentication systems and media credibility [3].

To counter these threats, audio deepfake detection has emerged as a critical research area in digital forensics. Traditional approaches rely on spectrogram-based analysis [4], which often struggles with unseen synthesis techniques or audio compression artifacts. Recent research explores **end-to-end learning from raw waveforms**, allowing models to automatically learn discriminative temporal and spectral features [5].

This paper presents a replication and evaluation of a **RawNetLite-based model** designed for end-to-end detection of fake and real audio. The model's simplicity and computational efficiency make it suitable for real-time applications. In addition to technical evaluation, this work reflects on the ethical implications of deepfake technologies, emphasizing the need for responsible AI deployment.

A. Related Work

Early detection methods relied heavily on handcrafted features such as Mel-Frequency Cepstral Coefficients (MFCCs) and spectrogram-based Convolutional Neural Networks (CNNs) [6]. Later works introduced systems like **RawNet2** [7], which process raw audio to capture both short-term texture and long-term rhythm. Models such as **AASIST** [8] and **RawGAT-ST** [9] improved generalization through attention and graph modules.

However, many of these models face challenges when applied to unseen conditions (cross-domain detection). Overfitting to dataset-specific features leads to poor generalization across different synthesis and recording environments.

Recent efforts therefore emphasize **lightweight architectures** and **data augmentation** for robustness [10]. In this context, the RawNetLite framework provides an efficient balance between accuracy and computational load, making it an ideal candidate for replication in educational and research settings

2.Literature Survey

The challenge of detecting audio deepfakes has become increasingly important with the emergence of highly realistic speech synthesis technologies such as WaveNet, Tacotron 2, and VALL-E [1], [2]. These systems generate speech that closely resembles real human voices, creating risks for misinformation, identity fraud, and impersonation attacks. To address these issues, researchers have explored both **feature-based** and **end-to-end deep learning** approaches for fake audio detection.

A. Early Feature-Based Methods

Initial detection systems relied on handcrafted acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs) and Constant-Q Cepstral Coefficients (CQCCs), which capture spectral characteristics of speech [4]. These features were typically classified using Support Vector Machines (SVMs) or shallow Convolutional Neural Networks (CNNs). Although computationally efficient, these systems suffered from limited robustness, particularly when tested on unseen datasets or varying recording conditions. The ASVspoof challenges [5] helped benchmark such methods and revealed 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** that operate directly on raw waveforms. This approach allows neural networks to automatically learn low- and high-level acoustic features that are critical for detecting synthesis artifacts. The **RawNet** and **RawNet2** architectures [6], [7] introduced convolutional-recurrent structures capable of learning both

spectral and temporal cues directly from the waveform. Models such as **AASIST** [8] and **RawGAT-ST** [9] further improved generalization by incorporating attention mechanisms and graph-based layers.

However, these advanced systems tend to be computationally expensive, limiting their practical deployment. The **RawNetLite** model addressed this gap by providing a **lightweight and efficient architecture** while maintaining strong detection accuracy. This study replicates and evaluates the RawNetLite framework as part of an educational and analytical project.

C. Cross-Dataset Generalization and Data Augmentation

A major challenge identified in the literature is **cross-domain generalization**—models trained on one dataset often fail when evaluated on others due to differences in recording devices, codecs, or synthesis methods. To mitigate this, researchers have explored **multi-dataset training**, **data augmentation**, and **specialized loss functions**.

Data augmentation techniques such as pitch shifting, time stretching, and additive noise were shown to improve model robustness [10]. Multi-domain training strategies, where models are trained on a combination of datasets (e.g., FakeOrReal, AVSpoof2021, and CodecFake), also enhance adaptability to unseen scenarios. In addition, **Focal Loss** [11] has been widely adopted to emphasize difficult examples during training, reducing overfitting and improving performance on minority classes.

D. Self-Supervised and Transformer-Based Methods

Recent progress in **self-supervised learning (SSL)** and **transformer architectures** has also influenced audio deepfake detection. Models such as **Wav2Vec 2.0** and **WavLM** have demonstrated that pretraining on massive unlabeled corpora can yield embeddings that generalize across spoofing types [12], [13]. These representations, when combined with lightweight classifiers, achieve state-of-the-art results on several benchmarks. However, SSL-based approaches require heavy computation and large datasets, making them less suitable for resource-constrained or real-time environments.

E. Ethical and Social Dimensions

In parallel with technical advances, scholars have highlighted the **ethical and societal implications** of deepfake technology. While synthetic speech can benefit accessibility and entertainment, malicious applications threaten individual privacy and public trust [14]. Responsible AI frameworks, such as the AI4People ethical guidelines [15], emphasize transparency, accountability, and informed consent in the deployment of generative technologies

3.Methodology

The proposed system for audio deepfake detection uses an end-to-end learning approach that works directly on raw audio waveforms. This section explains the datasets, preprocessing steps, and the RawNetLite model architectureused in this research.

3.1 Datasets Used

To train and evaluate the proposed model, three publicly available datasets were used — FakeOrReal, AVSpoof2021, and CodecFake. Each dataset provides unique challenges and helps in testing the model's ability to generalize across different audio sources and attack methods.

a) FakeOrReal (FoR)

This dataset contains both real and synthetic audio samples generated using advanced text-to-speech (TTS) and voice conversion (VC) systems. It serves as the main dataset for training and testing the model in a controlled environment.[SEP]A stratified split of 80% training, 10% validation, and 10% testing was applied to maintain class balance and avoid data overlap between phases.

b) AVSpoof2021

AVSpoof2021 includes diverse spoofing techniques such as replay attacks, TTS, and VC across different acoustic environments. It was mainly used for cross-dataset testing to evaluate the model's performance on unseen data. A balanced subset of real and fake samples was also used during training in domain-mix configurations.

c) CodecFake

This dataset focuses on real-world distortions like codec compression and transmission artifacts. It helps test how well the system performs when the audio quality is degraded. In some experiments, a subset of this dataset was used for training to improve the model's robustness under noisy or compressed conditions.

3.2 Preprocessing

All audio samples were:

- Converted to mono and resampled to 16 kHz, which balances quality and efficiency.
- Normalized to prevent amplitude variations from affecting learning.
- Trimmed or zero-padded to a fixed duration of 3 seconds (48,000 samples) for consistency.

Unlike traditional methods that use spectrograms or MFCCs, this model directly processes raw waveforms. This approach allows the network to automatically learn relevant patterns and avoids losing fine time-domain information.

3.3 Data Augmentation

To make the model more robust to real-world variations, several waveform-level augmentations were applied randomly during training:

- **Pitch Shift**: Simulates voice tone variation (±2 semitones).
- **Time Stretch**: Changes the speed slightly (0.9×–1.1×).
- **Gaussian Noise**: Adds low-level background noise (amplitude 0.001–0.015).

These augmentations mimic real conditions such as phone calls or online communication environments and help the model generalize better.

3.4 Model Architecture – RawNetLite

The RawNetLite model is a simplified and efficient version of the original RawNet. It learns directly from 1D audio waveforms and is suitable for real-time applications.[SEP]The architecture includes:

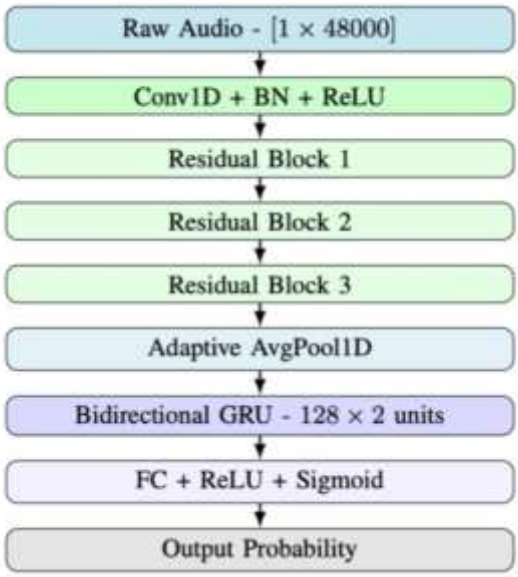

Fig. Architecture Of RawNetLite Model

1. 1D Convolutional Layer – Extracts low-level temporal features.
2. Three Residual Blocks – Capture hierarchical and local speech patterns.
3. Adaptive Average Pooling – Reduces the feature dimension.
4. Bidirectional GRU (128×2 units) – Learns long-term temporal dependencies like tone and rhythm.
5. Fully Connected Layers with Sigmoid Activation – Output a probability score (real or fake).

This structure allows RawNetLite to balance accuracy and computational cost, making it lightweight but powerful.

3.5 Training Setup

The model was trained using:

- **Adam optimizer** with a learning rate of 1×10^{-4},
- **Batch size** of 16, and
- **Early stopping** based on validation F1-score.

Training was carried out for up to 10 epochs. In augmented training, audio transformations were applied dynamically during each epoch.

3.6 Loss Functions

Two loss functions were explored:

a) Binary Cross-Entropy (BCE) Loss

Used as the baseline, it treats all samples equally and measures how well the model distinguishes between real and fake audio.

b) Focal Loss

Used to improve the focus on difficult or rare samples, especially in unbalanced datasets.[SEP]It reduces the influence of easily classified samples, forcing the model to learn from challenging cases.[SEP]Focal Loss was found to significantly improve cross-dataset accuracy.

3.7 Evaluation Metrics

To measure the performance of the system, the following metrics were used:

- Accuracy
- Precision
- Recall
- F1-Score
- Equal Error Rate (EER)

These metrics provide a balanced evaluation of how well the system detects fake speech while minimizing false alarms.

4.Result and Discussion

This section presents the experimental setup, evaluation protocols, and the results obtained for different datasets and training configurations. The proposed RawNetLite model was evaluated using FakeOrReal, AVSpoof2021, and CodecFake datasets to test both in-domain performance and cross-domain generalization.

4.1 Training Setup

All experiments were conducted using an NVIDIA RTX GPU and a multi-core processor. The model was trained using the Adam optimizer with a learning rate of 1×10^{-4}, a batch size of 16, and early stopping based on validation F1-score.[SEP]Training was carried out for up to 10 epochs.[SEP]During the augmented training phase, waveform-level transformations such as pitch

shifting, time stretching, and noise addition were applied randomly to make the system more robust to real-world distortions

4.2 Evaluation Protocol

To thoroughly analyze performance, three evaluation settings were considered:

1. **In-domain Testing**: Training and testing on the same dataset (FakeOrReal).
2. **Cross-domain Testing**: Training on one or two datasets and testing on unseen datasets (e.g., **CodecFake** or AVSpoof2021).
3. **Triple-domain Training**: Combining all three datasets to improve diversity and generalization.

Performance was measured using Accuracy, Precision, Recall, F1-score, and Equal Error Rate (EER).

4.3 In-Domain Performance

When trained and tested on the FakeOrReal dataset, the model achieved excellent results. It reached an F1-score of 99.2% and an EER of 0.29%, indicating that RawNetLite can effectively distinguish real and fake voices when both come from the same domain. This confirms that the proposed model architecture is capable of learning strong discriminative features directly from raw audio waveforms.

4.4 Cross-Domain Evaluation

Cross-dataset experiments were performed to test the generalization of the model to unseen conditions.

a) Testing on CodecFake

When evaluated on the CodecFake dataset, the baseline model achieved lower performance with an F1-score of only 17.8% and an EER around 50%. This drop in accuracy was mainly due to codec distortions and unseen data characteristics. However, when trained using domain-mix training (combining FakeOrReal and AVSpoof2021), performance improved slightly, showing the importance of exposing the model to diverse training data.

b) Testing on AVSpoof2021

In the AVSpoof2021 test set, the baseline model initially achieved 55.7% accuracy. After applying domain-mix training, the performance improved significantly with a fake F1-score of 82.4% and an EER of 16.6%. This result clearly shows that adding data from multiple domains enhances the model's adaptability to new spoofing techniques.

4.5 Effect of Focal Loss

Replacing Binary Cross-Entropy with Focal Loss led to major improvements, especially in cross-dataset detection. For example, on AVSpoof2021, fake recall increased from 44.9% to 71.9%, and the F1-score improved from 55.8% to 79.5%. This means Focal Loss helped the model focus more on hard-to-detect samples and reduced overfitting on easier ones. Importantly, it did not affect in-domain results, which stayed above 96% F1-score.

4.6 Triple-Domain Training

To further improve robustness, the model was trained using all three datasets—**FakeOrReal**, **AVSpoof2021**, and **CodecFake**. This **triple-domain** training achieved an F1-score of 78.6% on CodecFake and an overall accuracy of 83.9% across all test sets. The combined AVSpoof + CodecFake test set reached an F1-score of 83.4% and an EER of 16.4%, proving that domain diversity significantly strengthens generalization.

4.7 Effect of Audio Augmentation

Waveform-level audio augmentation was also tested as a low-cost alternative to using additional datasets. By applying pitch shifting, time stretching, and Gaussian noise during training, the model's cross-dataset performance improved noticeably.

The augmented model achieved:

- **FakeOrReal**: 97.4% F1-score
- **CodecFake**: 74.6% F1-score
- **AVSpoof2021**: 74.0% F1-score
- **Triple-Domain**: 80.8% F1-score

These results show that augmentation effectively simulates real-world variability and improves model resilience.

4.8 Discussion

The results highlight the main challenge of audio deepfake detection — the domain gap between training and testing data. While the model performs almost perfectly on known datasets, its performance drops when facing unseen attacks. Techniques such as domain-mix training, Focal Loss, and audio augmentation help bridge this gap, improving generalization and robustness.

Compared to larger or self-supervised models, RawNetLite achieves a strong accuracy-efficiency balance, making it suitable for real-world use where speed and computational cost matter.

5. CONCLUSIONS

This paper presented a replication and analysis of a lightweight end-to-end model for detecting audio deepfakes from raw waveforms. The RawNetLite architecture demonstrated high in-

domain accuracy and improved cross-domain generalization when trained with mixed datasets and Focal Loss.

Future research may explore **self-supervised pretraining**, **domain-adversarial learning**, and **open-set recognition** to enhance adaptability in unseen conditions. Additionally, collaborations between AI researchers, ethicists, and policymakers are essential to mitigate the risks associated with generative voice technologies

6.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

7.REFERENCES

[1] A. van den Oord *et al.*, "WaveNet: A generative model for raw audio," *arXiv preprint 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 International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2018.
[4] M. Todisco *et al.*, "ASVspoof 2019: Future directions for spoofed and fake audio detection," *IEEE Journal of Selected Topics in Signal Processing*, 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 Transactions on Information Forensics and 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, pp. 115–126, 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," *IEEE International Conference on Computer Vision (ICCV)*, 2017.
[12] ISO/IEC 30107-3, "Information technology—Biometric presentation attack detection—Part 3: Testing and reporting," *ISO Standard*, 2017.
[13] H. Westerlund, "The emergence of deepfake technology: A review," *Technology in Society*, vol. 67, 2021.
[14] L. Floridi *et al.*, "AI4People—An ethical framework for a good AI society," *Minds and Machines*, vol. 28, no. 4, 2018.