# Enhancing Deepfake Detection: A Study Using WavLM and Advanced RawBoost Augmentation Techniques

*Nhan Tri Do, Loi Nguyen Hoang, Phuong Ta Viet, Kien Phan Trung*

VinBigData Joint Stock Company, Vietnam

{dotrinhan99, hoangloi2001, tvphuong10, trungkien.it.98}@gmail.com

## Abstract

Automatic Speaker Verification (ASV) systems are increasingly vulnerable to sophisticated spoofing attacks, particularly those involving deepfake audio. This paper presents our approach to the challenges in Task 1 of the ASVSpoof 2024 competition, focusing on deepfake detection to classify utterances is spoof or bonafide. To enhance model robustness, we employed self-supervised learning (SSL) models, specifically fine-tuning WavLM for feature extraction due to its superior performance in noisy environments. We utilized RawBoost augmentation techniques to simulate real-world audio distortions. Experimental results demonstrate that our approach significantly improves detection accuracy, achieving an EER of 2.85% with WavLM and further reducing to 2.69% with a fusion of WavLM and Wav2Vec2 models.

**Index Terms**: Deepfake Detection, RawBoost, WavLM

## 1. Introduction

Automatic Speaker Verification (ASV) systems are critical for ensuring the security and reliability of voice authentication processes. However, these systems are increasingly vulnerable to spoofing attacks, particularly those involving deepfake audio, which can convincingly mimic the speech characteristics of a genuine user. Therefore, we conduct improvement studies and tests through the dataset and evaluation system of ASVSpoof 2024. This competition aims to advance the state-of-the-art in ASV spoofing countermeasures through rigorous evaluation and benchmarking. This paper focuses specifically on Task 1: Deepfake Detection.

The data utilized for this task includes metadata for each utterance, such as gender, attack type, and label, sourced from the training part of the Multilingual Librispeech dataset. The input data comprises FLAC audio files, and the goal is to classify each utterance as spoof or bonafide. Performance is evaluated using three primary metrics: the Minimum Detection Cost Function (minDCF), the Cost of Log-Likelihood Ratio (Cllr), and the Equal Error Rate (EER).

During our experiments, we encountered challenges with model convergence using the ASVSpoof 2024 training set alone. Consequently, we participated in the open track, combining data from previous years to enhance model training. The use of derived datasets and pre-trained models such as Libri-Light, MLS English, and MUSAN was prohibited. However, data from CommonVoice, previous ASVspoof editions (excluding VCTK), and LibriSpeech were permitted, allowing for a comprehensive and diverse training dataset.

In this paper, we present our approach to deepfake detection, detailing our data preparation, augmentation strategies, and model architectures. Our methodology leverages advanced techniques in self-supervised learning (SSL) and sophisticated model structures to achieve robust and accurate detection of deepfake audio. The experimental results demonstrate significant improvements over baseline models, highlighting the effectiveness of our proposed solutions in addressing the challenges of ASV spoofing.

## 2. Related Work

Spoofing audio can be detected through explicit features, as demonstrated by BTS-E [1], audio deepfake detection using breathing-talking-silence encoder. This method leverages natural human sounds, such as breathing, which are challenging to synthesize using text-to-speech technologies. BTS-E employs three simple Gaussian Mixture Models (GMM) for each class: breathing, talking, and silence.

As methods for generating human-like voices, both in terms of naturalness and intonation, continue to improve, new approaches are being developed to update and counteract the latest attack techniques. For instance, CodecFake[2] utilizes AASIST trained with custom codec data. Similarly, AI-Synthesized Voice Detection Using Neural Vocoder Artifacts[3] employs RawNet[4] as its backbone and constructs a new dataset using contemporary vocoders, incorporating both binary loss for bonafide and spoof classification and vocoder classification loss.

In addition to fusion models like the combination of ResNet18, LCNN9, and RawNet2 used by the top team T23 SpeechPro in 2021 [5], new approaches have been proposed. These include SE-Rawformer which leveraging positional-related local-global dependency [6], a transformer-based model that outperforms AASIST, and HM-Conformer[7], a Conformer-based system with hierarchical pooling and multi-level classification token aggregation methods, which achieved a 15.71% EER in ASVSpoof 2021 [8] without the need for model fusion.

Graph-based spectro-temporal dependency modeling introduced in 2023 [9], achieved an EER of 0.53%, outperforming AASIST's 0.83% EER on the logical access ASVspoof 2019 task [10]. This model segments spectral features into patches and constructs a graph where nodes are patch embeddings, and edge weights are computed using the dot-product of nodes. This approach is based on graph neural networks, similar to AASIST.

Beyond traditional feature extraction methods, a new direction involves using pre-trained semi-supervised learning models for speech feature extraction. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 [11], uses AASIST as its backbone and replaces the sinc-layer front-end with a wav2vec 2.0 model, it combines this with a self-attentive aggregation layer and adds nuisance variability on-the-fly to the

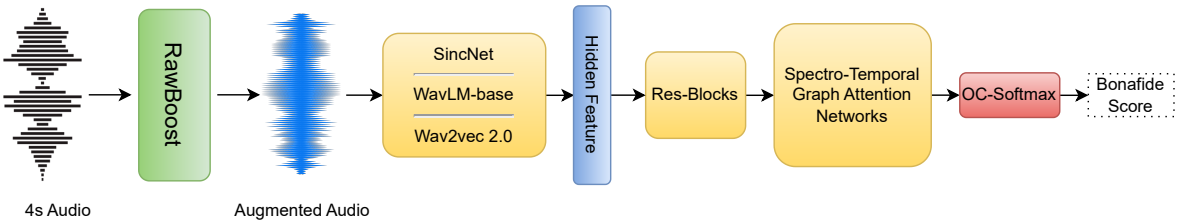


Figure 1: *Spectro-Temporal Graph Attention with WavLM Architecture Diagram*

| Feature Extractor | Backbone Model | EER | minDCF | actDCF | Cllr |
|---|---|---|---|---|---|
| Wav2Vec2 | Graph Attention | 4.26 | 0.1002 | 0.2255 | 0.3625 |
| Wav2Vec2 | Conformer | 4.49 | 0.1255 | 0.1676 | 0.6403 |
| WavLM | Graph Attention | 2.85 | 0.0816 | 0.1227 | 0.5552 |
| Fusion (Wav2Vec2 and WavLM) | Graph Attention | 2.69 | 0.0764 | 0.1622 | 0.2440 |

Table 1: *Experimental Results for Progress Phase*

existing training data.

To ensure models are robust and perform well across various domains, augmentation methods have been proposed, with RawBoost being a notable example. This method applies convolutive noise and impulsive signal-dependent additive noise strategies, which are particularly effective for the LA database as they simulate the convolutive and device-related noise sources characteristic of telephony applications. Additionally, stationary signal-independent additive noise, randomly colored, is used to counteract the effects of compression in the DF database. Furthermore, the Radian Weight Modification approach [12] in Self-Adaptive Continual Learning for Audio Deepfake Detection represents another significant contribution to the field.

# 3. Method

## 3.1. Self-supervised Feature Extractor

WavLM [13] is a versatile pre-trained model designed for robust speech processing, excelling in both clean and noisy environments for tasks like speech recognition, speaker identification, and emotion recognition. To enhance the generalization capability and robustness of our spoofing detection model, we utilized self-supervised learning (SSL) models to extract features from raw waveforms, opting for WavLM due to its superior performance in noisy environments, compared to wav2vec 2.0. During pretraining, the model processes raw audio input using a multi-layer convolutional feature encoder, transforming a sequence $\{x_t\}_{t=1}^{T}$ of $T$ time windows to output $\{z_t\}_{t=1}^{T}$.These representations are subsequently altered with noise and overlapping effects before being masked and passed into the Transformer encoder, which produces a sequence of hidden states $\{h^l\}_{l=1}^{L}$, where $L$ denotes the number of Transformer layers. Moreover, the model integrates gated relative position bias, which improves its capacity to attend to pertinent speech features effectively. WavLM is trained using a masked speech denoising and prediction task. This approach inherently captures speaker and speech-related features, as the training objective involves predicting pseudo-labels for the masked portions of the original speech.

Since the evaluation set contains audio used to train WavLM large, we conducted our experiments fairly by only evaluating with WavLM-base. We fine-tuned the WavLM base model using the LibriSpeech dataset, averaging the outputs of its layers to create feature vectors with a dimensionality of 768 as in Figure 2.

## 3.2. Spectrotemporal Graph Attention Network

Building on the AASIST framework, we replaced the original sinc-layer front-end with the WavLM model. This backbone spectrotemporal graph attention network takes the hidden features input extracted from the pretrained WavLM model. These features are passed through a linear post-processing layer to reduce the feature dimension before being fed into RawNet2 [14], which consists of 6 residual blocks. This structure enables learning high-level features that represent of channels, spectral, and time frames. Spectral and temporal representations are then created using max pooling functions and processed through a graph attention network. These representations are combined into a heterogeneous spectrotemporal graph using heterogeneous stacking graph attention layer - HSGAL. The nodes of this graph are further fed into two parallel HS-GAL modules to learn spoofing features before merging into a final graph. A set of operations known as Readout is performed on nodes of this graph, including node-wise maximum and average for the spectral and temporal nodes, respectively. Each operation produces a 32-dimensional feature, which then concatenated and results in a 160-dimensional vector. This vector is then passed through a fully connected layer to produce the two classes: bonafide and spoof. Additionally, we also experimented with replacing Graph HS-GAL with a Conformer [15] and integrated a Retention Network [16] to enhance model complexity and improve inference speed.

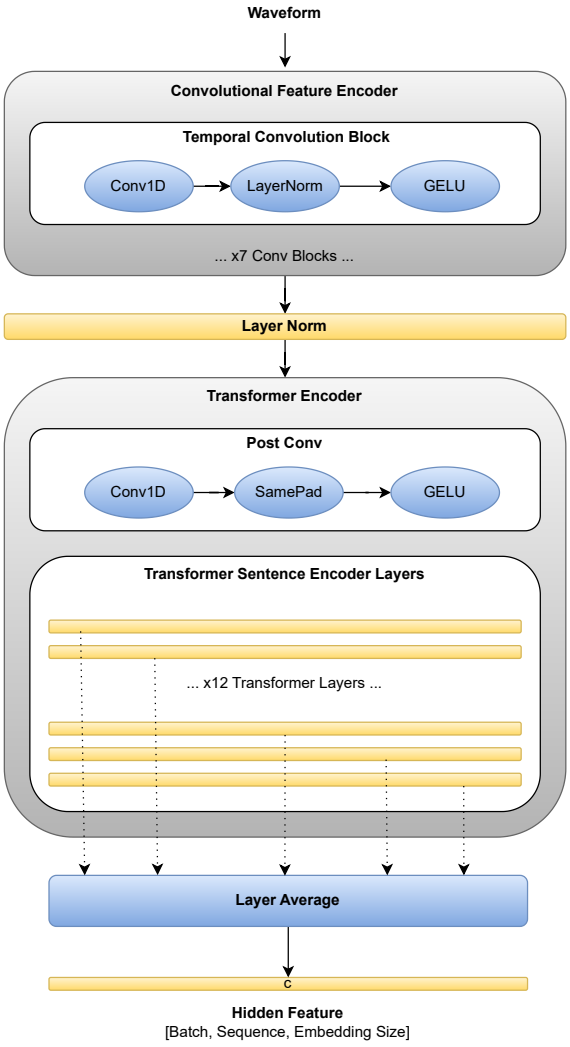


Figure 2: *Pretrained WavLM Based Feature Extractor*

### 3.3. Loss for AntiSpoofing

Our loss function strategy included a weighted cross-entropy loss to address class imbalance between the Bonafide and Spoof classes, alongside OCSoftmax loss configured to focus on detecting bonafide cases more accurately [17]. In practice, new voice attack methods are constantly being developed, leading to the emergence of unknown attacks. If the original Softmax loss for binary classification is used, the model may overfit to the attack methods present in the training set. OCSoftmax (One-Class Softmax) is a specialized variant of the standard softmax function that focuses on detecting bonafide audio and isolating spoofing attacks.

$$L_{\text{WCE}} = -\frac{1}{N}\sum_{n=1}^{N}\left[w_1 \cdot y_n \cdot \log(p_{n,1}) + w_0 \cdot (1-y_n) \cdot \log(p_{n,0})\right]$$

$$p_{n,c} = \frac{\exp(x_{n,c})}{\exp(x_{n,0}) + \exp(x_{n,1})} \quad \text{for } c \in \{0,1\}$$

$$L_{\text{OCS}} = \frac{1}{N}\sum_{i=1}^{N}\log\left(1 + e^{\alpha(m_{y_i} - \hat{w}_0 \cdot \hat{x}_i)(-1)^{y_i}}\right)$$

Despite exploring additional methods like Gradient Reversal Layers [18], EfficientNet Attention [19], Spec-ResNet, and various fusion techniques, the SSL-based approach consistently yielded the best results, demonstrating the effectiveness of combining advanced feature extraction and model architecture enhancements in detecting audio spoofing.

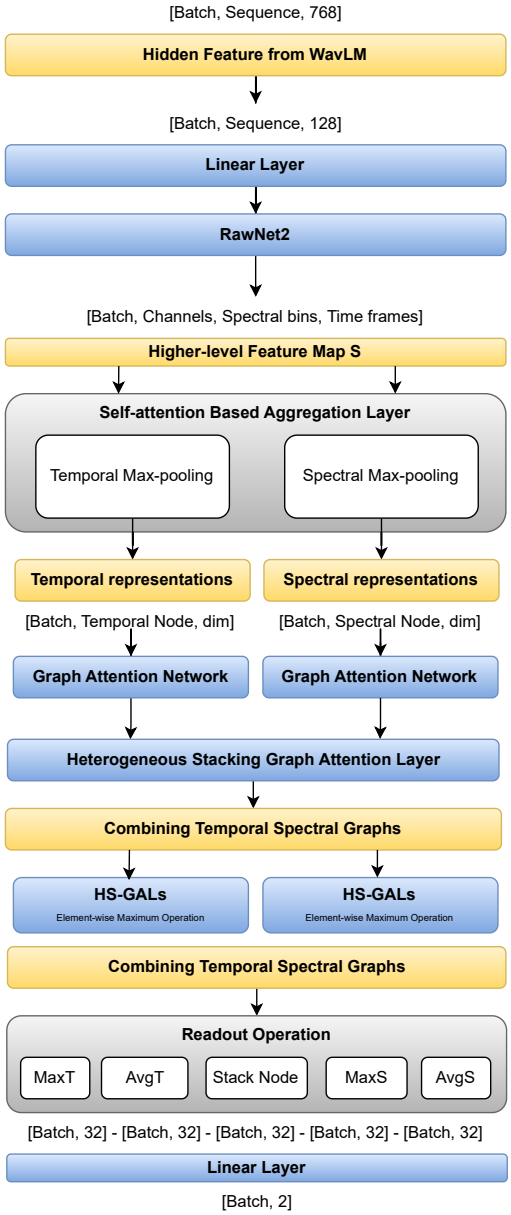


Figure 3: *AASIST Backbone - Spectrotemporal Graph Attention Network*

## 4. Experiment

We trained our models using 4-second audio segments, preprocessed with random padding and silence trimming, and augmented with RawBoost, which includes linear and non-linear convolutive noise, impulsive signal-dependent additive noise, and stationary signal-independent additive noise.

For the weighted cross-entropy loss, the weights assigned to bonafide and spoof classes are 0.9 and 0.1, respectively. For the OCSoftmax Loss, the scaling factor $\alpha$ is set to 20, and the margins for real and fake audio are set to 0.9 and 0.2, respectively, as suggested in the original paper.

The experiment was conducted with a batch size of 24, utilizing distributed training across two GeForce RTX 4090 GPUs and 10 CPU cores.

To evaluate and compare the model's performance, we focus on the EER metric, which is the point where the system's False Acceptance Rate (incorrectly accepting spoofed audio) and False Rejection Rate (incorrectly rejecting real audio) are equal. A lower EER indicates higher model accuracy. The experiments results are summarized in the following table, the WavLM with spectro-temporal graph attention and OCSoftmax achieved a significantly lower EER of 2.85%. Additionally, the fusion of WavLM and Wav2Vec2 models, weighted at 70% and 30% respectively, further reduced the EER to 2.69%, demonstrating the effectiveness of combining robust feature extraction with advanced model architectures and augmentation techniques.

Additionally, experiments using Conformer instead of the graph attention network indicated that it did not improve the model's ability to detect spoofing.

## 5. Conclusion

In this study, we addressed the challenge of deepfake audio detection in ASV systems, as part of the ASVSpoof 2024 competition. By leveraging advanced self-supervised learning models and robust augmentation techniques, we developed a detection framework that significantly enhances performance. The fine-tuning of WavLM for feature extraction, combined with Spectro-Temporal Graph Attention Networks, proved highly effective in improving model robustness and accuracy. Our experimental results, demonstrating an EER of 2.85% with WavLM and 2.69% with model fusion, highlight the potential of our approach in real-world applications. Future work will focus on optimizing model architecture with different hyper-parameters and exploring additional augmentation strategies.

## 6. Acknowledgements

We would like to express our sincere gratitude to the organizers of the ASVSpoof 5 competition for providing a platform to advance the research in automatic speaker verification and spoofing detection.

## 7. References

[1] T.-P. Doan, L. Nguyen-Vu, S. Jung, and K. Hong, "Bts-e: Audio deepfake detection using breathing-talking-silence encoder," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[2] H. Wu, Y. Tseng, and H.-y. Lee, "Codecfake: Enhancing anti-spoofing models against deepfake audios from codec-based speech synthesis systems," *arXiv preprint arXiv:2406.07237*, 2024.

[3] C. Sun, S. Jia, S. Hou, and S. Lyu, "Ai-synthesized voice detection using neural vocoder artifacts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 904–912.

[4] J.-w. Jung, H.-S. Heo, J.-h. Kim, H.-j. Shim, and H.-J. Yu, "Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification," *arXiv preprint arXiv:1904.08104*, 2019.

[5] A. Tomilov, A. Svishchev, M. Volkova, A. Chirkovskiy, A. Kondratev, and G. Lavrentyeva, "Stc antispoofing systems for the asvspoof2021 challenge," in *Proc. ASVspoof 2021 Workshop*, 2021, pp. 61–67.

[6] X. Liu, M. Liu, L. Wang, K. A. Lee, H. Zhang, and J. Dang, "Leveraging positional-related local-global dependency for synthetic speech detection," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[7] H.-s. Shin, J. Heo, J.-h. Kim, C.-y. Lim, W. Kim, and H.-J. Yu, "Hm-conformer: A conformer-based audio deepfake detection system with hierarchical pooling and multi-level classification token aggregation methods," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 581–10 585.

[8] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch *et al.*, "Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2507–2522, 2023.

[9] X. Zhang *et al.*, "Graph-based spectro-temporal dependency modeling for anti-spoofing," arXiv, 2023. [Online]. Available: https://arxiv.org/pdf/2304.13085

[10] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee *et al.*, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101114, 2020.

[11] H. Tak, M. Todisco, X. Wang, J.-w. Jung, J. Yamagishi, and N. Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," *arXiv preprint arXiv:2202.12233*, 2022.

[12] X. Zhang, J. Yi, C. Wang, C. Y. Zhang, S. Zeng, and J. Tao, "What to remember: Self-adaptive continual learning for audio deepfake detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, 2024, pp. 19 569–19 577.

[13] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *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.

[14] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with rawnet2," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6369–6373.

[15] E. Roselló Casado, A. Gómez Alanís, Á. M. Gómez García, A. M. Peinado Herreros *et al.*, "A conformer-based classifier for variable-length utterance processing in anti-spoofing," 2023.

[16] Y. Sun, L. Dong, S. Huang, S. Ma, Y. Xia, J. Xue, J. Wang, and F. Wei, "Retentive network: A successor to transformer for large language models," *arXiv preprint arXiv:2307.08621*, 2023.

[17] Y. Zhang, F. Jiang, and Z. Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Processing Letters*, vol. 28, pp. 937–941, 2021.

[18] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," 2015.

[19] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.