International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(16s) 2026 pp. 1458-1470
https://doi.org/10.70917/ijcisim-2026-4653
Article

A Hybrid MFCC–WavLM Feature Fusion for Audio Deepfake Detection

K. Sunil Kumar[1], Madduluri Suneetha[2], K. R. Anudeep Laxmi Kanth[3], M. Manoj Kumar[4], T. Kishore Kumar[5]

[1,3,4,5]Department of ECE, National Institute of Technology Warangal, Warangal, Telangana, India
[2]School of Electronics Engineering (SENSE), VIT-AP University, Andhra Pradesh, India
[1]sunil.veena10@gmail.com, 3kralk1989@gmail.com, 4myakalamanojkumar@gmail.com , 5kishoret@nitw.ac.in
[2]madduluri.suneetha@vitap.ac.in

Abstract: Recent advances in generative artificial intelligence have enabled highly realistic speech synthesis using text-to-speech (TTS), voice conversion (VC), and neural voice cloning techniques, posing significant security threats to Automatic Speaker Verification (ASV) systems. Conventional handcrafted features such as Mel-Frequency Cepstral Coefficients (MFCCs) effectively capture spectral characteristics but may fail to detect sophisticated artifacts introduced by modern speech synthesis models. In contrast, self-supervised models such as WavLM provide rich contextual representations but may not fully capture complementary lowlevel acoustic information. This paper proposes a hybrid audio deepfake detection framework that combines an 80-dimensional MFCC feature vector with a 768-dimensional WavLM embedding to form an 848-dimensional hybrid representation. The extracted features are standardized using StandardScaler, balanced using the Synthetic Minority Over-sampling Technique (SMOTE), and classified using a Deep Neural Network (DNN). Experiments on the ASVspoof2019 Logical Access (LA) dataset show that the proposed framework achieves 94.8% classification accuracy and a weighted F1-score of 0.95 on a 5,000-sample evaluation subset, demonstrating the effectiveness of the hybrid representation for spoofed speech detection.
Keywords: Audio Deepfake Detection, Automatic Speaker Verification, ASVspoof2019, MFCC, WavLM, Self-Supervised Learning, Deep Neural Network, Speech Forensics

1. Introduction

Recent advances in generative artificial intelligence have enabled highly realistic synthetic speech through text-to-speech (TTS), voice conversion (VC), and neural voice cloning technologies. Although these technologies have enabled applications in virtual assistants, speech restoration, and human–computer interaction, they also introduce security risks for Automatic Speaker Verification (ASV) systems. AI-generated speech can be used to impersonate individuals and deceive speaker verification systems, increasing the need for reliable audio spoof and deepfake detection methods.

Traditional audio spoof detection systems commonly rely on handcrafted acoustic features such as Mel-Frequency Cepstral Coefficients (MFCC), Constant-Q Cepstral Coefficients (CQCC), and Linear Frequency Cepstral Coefficients (LFCC). Among these, MFCCs provide a compact representation of the spectral characteristics of speech with relatively low computational complexity. However, handcrafted features primarily describe low-level acoustic properties and may not adequately represent the contextual and temporal characteristics associated with sophisticated synthetic speech generation.

Recent advances in self-supervised learning (SSL) have introduced powerful speech representation models such as wav2vec 2.0, HuBERT, and WavLM. These models learn contextual speech representations from large-scale unlabeled speech data and have demonstrated strong performance across a range of speech processing tasks. In

particular, WavLM provides high-level contextual representations that can capture information beyond conventional handcrafted acoustic descriptors. However, relying exclusively on SSL-based representations may not fully exploit low-level spectral characteristics that can also provide useful cues for distinguishing bona fide and spoofed speech.

These observations motivate the integration of complementary handcrafted and self-supervised speech representations for audio deepfake detection. In this work, MFCC features are combined with pretrained WavLM embeddings to jointly represent low-level spectral and high-level contextual characteristics of speech. Specifically, 40 MFCC coefficients are summarized using their mean and standard deviation to obtain an 80-dimensional acoustic representation, while mean pooling of WavLMBase hidden representations produces a 768-dimensional contextual embedding. These features are concatenated to form an 848-dimensional hybrid representation, which is subsequently standardized and used to train a Deep Neural Network (DNN) classifier.

The proposed framework is evaluated using the ASVspoof2019 Logical Access (LA) dataset. To address the class imbalance present in the training data, Synthetic Minority Over-sampling Technique (SMOTE) is applied to the training feature space. The resulting framework combines complementary acoustic and contextual information while maintaining a relatively compact classification architecture.

The major contributions of this work are summarized as follows:

- A hybrid MFCC–WavLM feature representation is developed for audio deepfake detection by combining handcrafted spectral descriptors with pretrained self-supervised speech representations.
- An 848-dimensional feature representation is constructed by integrating an 80-dimensional MFCC representation with a 768-dimensional WavLM embedding.
- A DNN-based classification framework is developed using standardized and SMOTE-balanced training features for bona fide and spoof speech classification.
- Experimental evaluation on the ASVspoof2019 LA dataset is performed to assess the effectiveness of the proposed hybrid representation.

2. Related Work

Audio deepfake and speech spoofing detection has evolved from conventional handcrafted acoustic features toward deep learning and self-supervised speech representation models. The ASVspoof challenge series has played an important role in this development by providing benchmark datasets and standardized evaluation protocols for assessing countermeasures against synthetic, converted, and replayed speech [2], [17].

A. Conventional Acoustic Features for Spoof Detection Early speech spoofing detection systems primarily relied on hand-crafted acoustic features, including Mel-Frequency Cepstral Coefficients (MFCC), Constant-Q Cepstral Coefficients (CQCC), and Linear Frequency Cepstral Coefficients (LFCC), combined with conventional machine learning classifiers such as Gaussian Mixture Models (GMMs), Support Vector Machines (SVMs), and Random Forests [1], [10], [11]. MFCCs provide a compact representation of the spectral characteristics of speech and have been widely adopted because of their relatively low computational complexity [1]. CQCCs were introduced as an effective acoustic representation for spoofing countermeasures, while LFCCs have also been investigated for synthetic speech detection [10], [11]. However, handcrafted features primarily describe low-level acoustic characteristics and may have limited ability to represent the complex temporal and contextual artifacts introduced by modern neural speech synthesis systems [11], [14], [18].

B. Deep Learning-Based Audio Spoof Detection To overcome the limitations of manually designed features, deep learning-based approaches have increasingly been investigated for audio spoof detection. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based architectures have demonstrated the ability to learn discriminative representations from speech signals and acoustic features [12], [14]. CNN-based approaches can capture local spectral patterns, while recurrent and Transformer-based architectures can model

temporal dependencies and contextual relationships within speech representations [12], [14]. Although these approaches can improve spoof detection performance, increasingly complex architectures may introduce additional computational and memory requirements, which can be challenging for resourceconstrained deployment scenarios [14], [19].

C. Self-Supervised Speech Representations Self-supervised learning (SSL) has significantly advanced speech representationlearning by enabling models to learn general-purpose representations from large-scale unlabeled speech corpora. Models such as wav2vec 2.0, HuBERT, and WavLM have demonstrated strong representation-learning capabilities across a variety of speech processing tasks [3]–[5], [20]. In particular, WavLM is designed to learn contextual speech representations and has demonstrated effectiveness in capturing speaker-related and speech-content information [3]. SSL-based representations have also been investigated for audio spoofing detection and have shown advantages over conventional handcrafted acoustic representations in capturing higher-level speech characteristics [14], [15], [20]. Despite these advantages, SSL representations are generally high-dimensional and primarily emphasize learned contextual representations. Consequently, relying exclusively on SSL embeddings may not fully exploit complementary low-level spectral characteristics that can provide useful information for distinguishing bona fide and spoofed speech [14], [15], [20].

D. Hybrid Feature Representation and Fusion Approaches The complementary nature of handcrafted acoustic featuresand learned speech representations has motivated the development of feature-fusion approaches for spoof detection. Such approaches aim to combine different levels of speech information to improve the discriminative capability of the detection system. For example, the PACA-MTransformer framework incorporates presentation-aware contextual attention and a modified Transformer architecture to jointly exploit handcrafted acoustic features and contextual information for audio deepfake detection [25]. Although such Transformer-based approaches can effectively model contextual information, their increased architectural complexity may result in higher computational and inference costs, which can limit their suitability for resource-constrained deployment [14], [19]. Other studies have also demonstrated the effectiveness of combining learned speech representations with complementary acoustic information for spoof detection [15], [20]. These findings indicate that integrating multiple levels of speech representation can provide more discriminative information than relying on a single feature type.

E. Research Gap and Motivation From the existing literature, handcrafted features such as MFCC provide compact low-levelspectral information [1], [10], [11], whereas SSL models such as WavLM provide richer contextual speech representations [3], [15], [20]. Existing studies have investigated these representations independently as well as within relatively complex deep learning architectures [14], [15], [25]. However, there remains scope for investigating a simpler feature-level fusion strategy that combines handcrafted spectral descriptors with pretrained contextual speech representations while maintaining

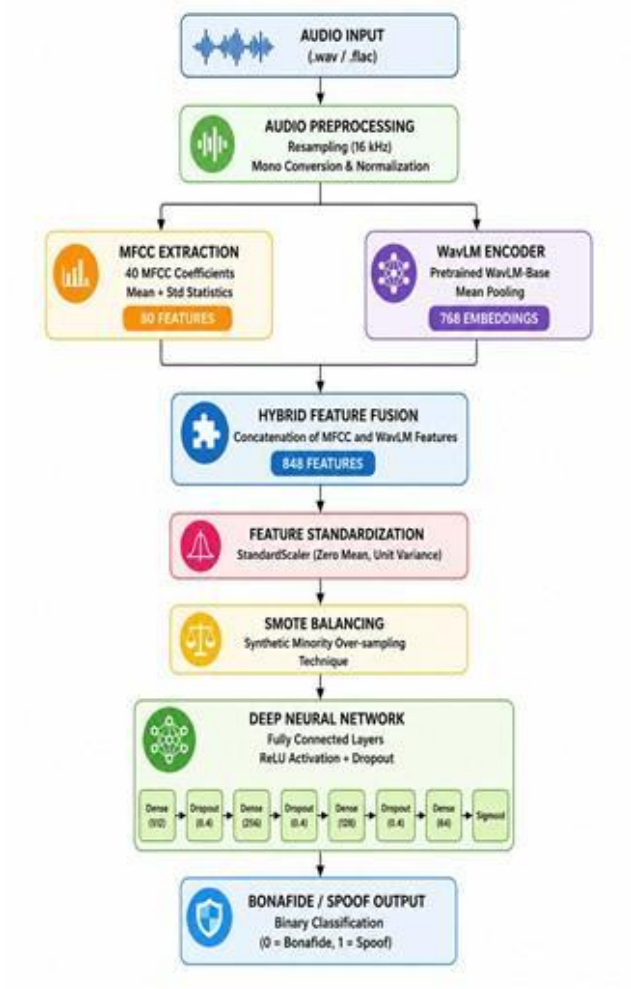

Fig. 1: Overall workflow of the proposed hybrid MFCC–WavLM audio deepfake detection framework.

a comparatively compact classification architecture.Motivated by this gap, the present work combines MFCC features with pretrained WavLM embeddings at the feature level. The proposed representation integrates low-level spectral characteristics with high-level contextual information and is subsequently classified using a Deep Neural Network (DNN). This design aims to exploit the complementary characteristics of both feature types while avoiding the architectural complexity associated with large end-to-end Transformer-based detection models [14], [19], [25].

3. Proposed Methodology

The proposed hybrid audio deepfake detection framework consists of five main stages: audio preprocessing, feature extraction, hybrid feature fusion, feature preprocessing, and classification.As illustrated in Fig. 1 1, the input speech signal is first preprocessed and subsequently passed through two parallel feature-extraction branches. The first branch extracts handcrafted MFCC features, while the second branch uses a pretrained WavLM model to obtain contextual speech embeddings. The resulting features are fused to form an 848-dimensional representation, followed by feature standardization, class balancing using SMOTE, and DNN-based binary classification.

A. Dataset Description

The proposed framework is evaluated using the ASVspoof2019 Logical Access (LA) dataset, which contains bona fide and spoofed speech generated using various text-to-speech (TTS) and voice conversion (VC) systems [2],

[17]. The corpus is divided into training, development, and evaluation subsets. The training and development subsets are used for model development, while the evaluation partition is used for final performance assessment.

All speech recordings are stored in FLAC format with a sampling frequency of 16 kHz. The diversity of spoofing algorithms included in the dataset makes it suitable for evaluating the robustness and generalization capability of deepfake detection systems.

TABLE I: ASVspoof2019 LA Dataset Statistics

Subset	Bona fide	Spoof	Total
Training	2,580	22,800	25,380
Development	2,548	22,296	24,844
Evaluation	7,355	63,882	71,237

B. Audio Preprocessing

Each speech utterance is resampled to 16 kHz and converted to mono to obtain a uniform input representation. The waveform amplitude is normalized before feature extraction. Since the utterances have different durations, fixed-length representations are subsequently obtained using mean and standard-deviation pooling for MFCC features and temporal mean pooling for WavLM embeddings.

C. MFCC Feature Extraction

Mel-Frequency Cepstral Coefficients (MFCC) are extracted to capture the perceptually relevant spectral characteristics of speech. For each utterance, 40 MFCC coefficients are computed using the Librosa library. The mean and standard deviation of each coefficient across all frames are calculated to obtain a fixed-length feature vector.

The resulting MFCC representation is expressed as

$$\mathbf{M} = [\mu 1, \mu 2, ..., \mu 40, \sigma 1, \sigma 2, ..., \sigma 40] \quad (1)$$

where μi and σi denote the mean and standard deviation of the *ith* MFCC coefficient, respectively. This produces an 80-dimensional handcrafted acoustic feature vector.

D. WavLM Embedding Extraction

The pretrained WavLM-Base model (microsoft/wavlm-base) is used as a self-supervised speech representation extractor. The normalized waveform is provided to WavLM, which generates frame-level hidden representations. The pretrained WavLM parameters are kept fixed during DNN training, and temporal mean pooling is applied to the last hidden-state representation to obtain a fixed-length 768-dimensional utterance-level embedding.

Mean pooling is applied over the temporal dimension to obtain a fixed-length embedding represented as

$$\mathbf{W} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{h}_t. \quad (2)$$

where $\mathbf{h}t$ denotes the hidden representation at frame t, and T is the total number of frames. The resulting feature vector has a dimensionality of 768.

E. Hybrid Feature Fusion

The 80-dimensional MFCC representation and 768-dimensional WavLM embedding are concatenated along the feature dimension to construct the hybrid representation:

$$\mathbf{H} = [\mathbf{M};\mathbf{W}], \quad (3)$$

where $\mathbf{M} \in \mathbb{R}^{80}$ and $\mathbf{W} \in \mathbb{R}^{768}$. Consequently, the hybrid feature vector has a total dimensionality of 848. Therefore, the resulting hybrid feature vector has 848 dimensions. This representation jointly captures low-level spectral characteristics and higher-level contextual speech information.

F. Feature Standardization and SMOTE

The fused 848-dimensional feature vectors are standardized using StandardScaler. The scaler is fitted using the training data and subsequently applied to the corresponding evaluation features. To address the substantial class imbalance between bona fide and spoofed speech in the training data, SMOTE is applied to the training feature space to generate synthetic minority-class samples.

G. Deep Neural Network Classifier

The fused 848-dimensional feature vector is classified using a fully connected DNN. The network consists of four hidden layers with 512, 256, 128, and 64 neurons, respectively. ReLU activation is used in the hidden layers, while dropout regularization is employed to reduce overfitting. L2 regularization is applied to the first three dense layers. The final layer contains a single neuron with sigmoid activation to produce the probability of the input utterance being spoofed. The network is trained using the Adam optimizer with binary cross-entropy loss.

4. Experimental Setup

This section describes the experimental configuration adopted to evaluate the proposed hybrid MFCC–WavLM audio deepfake detection framework. All experiments were conducted using the TensorFlow deep learning framework on the ASVspoof2019 Logical Access (LA) dataset. The proposed system was trained using hybrid features obtained by concatenating handcrafted MFCC descriptors with contextual WavLM embeddings. The extracted features were standardized using StandardScaler, and class imbalance in the training set was addressed using the Synthetic Minority Over-sampling Technique (SMOTE).

A. Dataset

The experiments were conducted using the ASVspoof2019 Logical Access (LA) dataset [2], [17]. The training partition contains 25,380 utterances, including 2,580 bona fide and 22,800 spoofed samples, while the development partition contains 24,844 utterances. The complete evaluation partition consists of 71,237 utterances, including 7,355 bona fide and 63,882 spoofed samples.

For the final evaluation, a subset of 5,000 utterances was randomly selected from the ASVspoof2019 LA evaluation partition using a fixed random seed of 42. The selected subset contains 530 bona fide and 4,470 spoofed utterances. This evaluation subset was kept separate from the training data and was used only for final performance assessment.

B. Feature Extraction

The audio signals were processed at a sampling rate of 16 kHz and converted to mono. MFCC features were extracted using 40 coefficients with a 25-ms analysis window and a 10-ms hop length. For each utterance, the mean and standard deviation of the MFCC coefficients were calculated, resulting in an 80-dimensional feature vector. For contextual representation, the pretrained WavLM-Base model was used as a fixed feature extractor. The frame-level hidden representations were mean-pooled along the temporal dimension to obtain a 768-dimensional utterance-level embedding. The MFCC and WavLM representations were concatenated to form an 848-dimensional hybrid feature vector.

C. DNN Training Configuration

The hybrid 848-dimensional features were classified using a fully connected DNN. The network consists of four hidden layers containing 512, 256, 128, and 64 neurons, respectively. ReLU activation was used in the hidden layers, with dropout regularization of 0.4, 0.4, and 0.3 applied to the first three hidden layers. L2 regularization with a coefficient of 10^{-4} was applied to the first three dense layers. The final output layer contains a single neuron with sigmoid activation for binary classification. The network was trained using the Adam optimizer and binary cross-entropy loss with a batch size of 64 for a maximum of 30 epochs. Early stopping with a patience of 5 epochs was employed based on validation loss, and the model weights corresponding to the best validation performance were restored.

D. Software and Implementation

The proposed framework was implemented using Python and TensorFlow, with Librosa used for MFCC extraction and the Hugging Face Transformers implementation used for WavLM feature extraction. Scikit-learn was used for feature standardization, data splitting, and performance evaluation, while the imbalanced-learn library was used for SMOTE-based class balancing.

E. Evaluation Metrics

The effectiveness of the proposed framework was assessed using standard classification metrics, including Accuracy, Precision, Recall, and F1-score. These metrics are defined as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Precision = \frac{TP}{TP + FP}, \tag{5}$$

$$Recall = \frac{TP}{TP + FN}, \tag{6}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \tag{7}$$

where *TP*, *TN*, *FP*, and *FN* denote the number of true positives, true negatives, false positives, and false negatives, respectively. The experimental parameters used for feature extraction, feature preprocessing, DNN training, and evaluation are TABLE II: Experimental Configuration of the Proposed MFCC–WavLM-DNN Framework

Parameter	Configuration
Dataset	ASVspoof2019 LA
Sampling Rate	16 kHz
Audio Format	Mono
MFCC Coefficients	40
MFCC Window Length	25 ms
MFCC Hop Length	10 ms
MFCC Feature Dimension	80
WavLM Model	WavLM-Base

WavLM Feature Dimension	768
Hybrid Feature Dimension	848
Feature Scaling	StandardScaler
Class Balancing	SMOTE
DNN Architecture	848–512–256–128–64–1
Hidden Layer Activation	ReLU
Dropout 0.4, 0.4, 0.3 L2 Regularization 10^{-4}	
Output Activation	Sigmoid
Optimizer	Adam
Loss Function	Binary Cross-Entropy
Batch Size	64
Maximum Epochs	30
Early Stopping Patience	5
Evaluation Samples	5,000
Random Seed	42

TABLE III: Classification Performance of the Proposed MFCC–WavLM-DNN Model

Class	Precision	Recall	F1-Score	Support
Bona fide	0.69	0.94	0.79	530
Spoof	0.99	0.95	0.97	4,470
Accuracy	–	–	94.8%	5,000
Macro Average	0.84	0.94	0.88	5,000
Weighted Average	0.96	0.95	0.95	5,000

summarized in Table II. The configuration was kept consistent with the implementation used for the proposed MFCC–WavLMDNN framework. The configuration in Table II defines the complete feature-processing and classification pipeline. In particular, the 80-dimensional MFCC representation and 768-dimensional WavLM embedding are concatenated to form an 848dimensional input to the DNN classifier.

5. Results and Discussion

A. Main Classification Results

The proposed MFCC–WavLM feature fusion framework was evaluated on a randomly selected subset of 5,000 utterances from the ASVspoof2019 LA evaluation partition. The selected subset consisted of 530 bona fide and 4,470 spoof utterances.

As shown in Table III, the proposed model achieved an overall classification accuracy of 94.8% and a weighted F1-score of

0.95.

For the bona fide class, the model achieved a precision of 0.69, recall of 0.94, and F1-score of 0.79. For the spoof class, the corresponding values were 0.99, 0.95, and 0.97, respectively. The higher F1-score obtained for the spoof class indicates that the proposed framework is particularly effective in identifying spoofed speech. However,

the comparatively lower precision for the bona fide class indicates that some spoofed utterances are classified as bona fide, highlighting an area for further improvement.

The overall weighted F1-score of 0.95 demonstrates that the proposed hybrid representation provides effective discrimination between bona fide and spoofed speech despite the class imbalance in the evaluation subset. The classification performance of the proposed MFCC–WavLM-DNN framework was evaluated on a randomly selected subset of 5,000 utterances from the ASVspoof2019 LA evaluation partition. The subset contains 530 bona fide and 4,470 spoof utterances. The class-wise precision, recall, F1-score, and support values are reported in Table III. As shown in Table III, the proposed framework achieves an overall accuracy of 94.8% and a weighted F1-score of 0.95. The spoof class achieves an F1-score of 0.97, whereas the bona fide class achieves an F1-score of 0.79. These results indicate that the proposed hybrid representation provides effective discrimination between bona fide and spoofed speech, although bona fide classification remains comparatively more challenging.

B. Confusion Matrix

The confusion matrix is used to examine the class-wise prediction behavior of the proposed MFCC–WavLM-DNN model. It provides a detailed view of correctly and incorrectly classified bona fide and spoofed utterances.

The confusion matrix shows that the proposed model correctly identifies the majority of both bona fide and spoofed speech samples. The model achieves a higher number of correct predictions for the spoof class, which is consistent with the higher

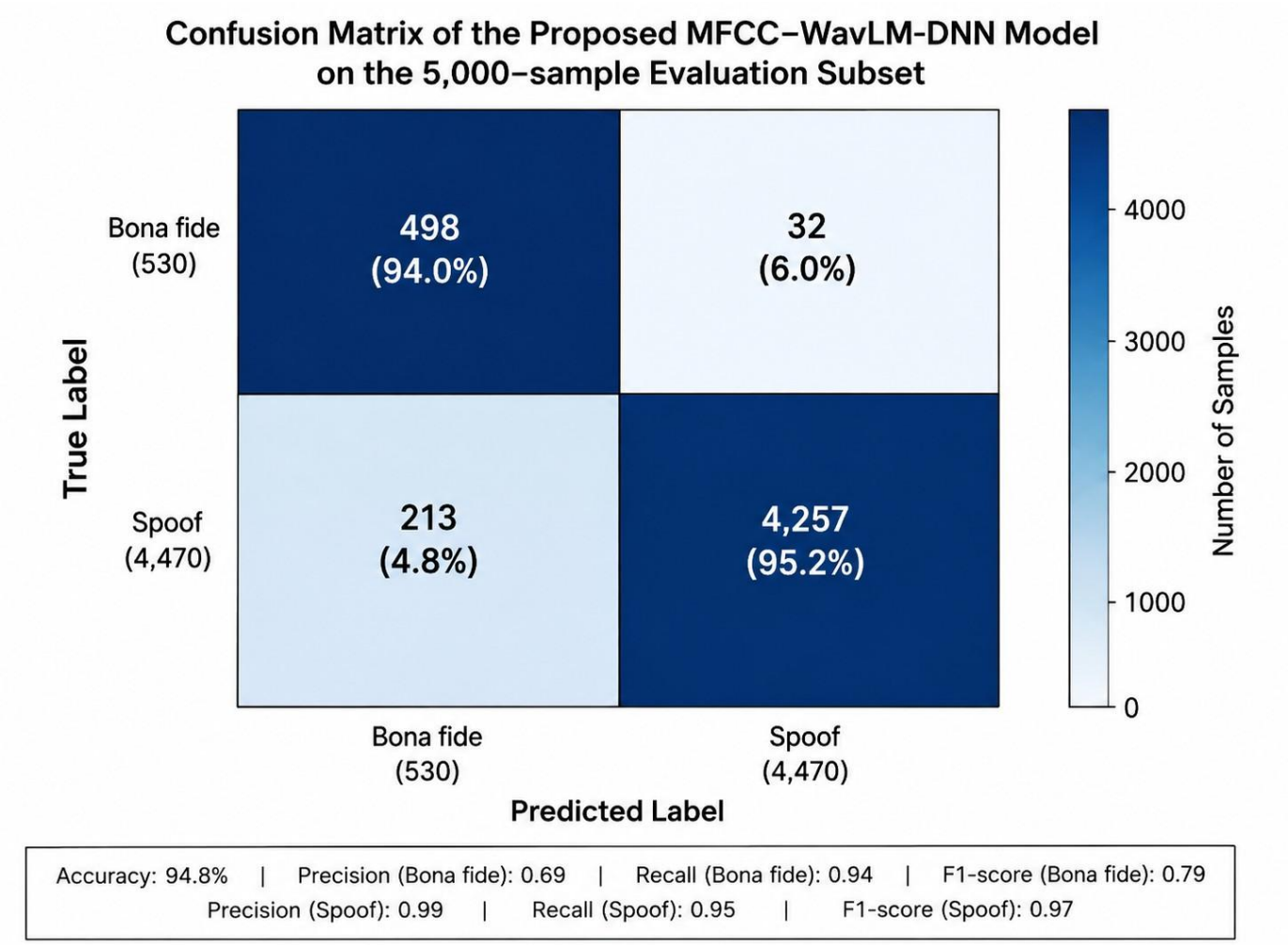

Fig. 2: Confusion matrix of the proposed MFCC–WavLM-DNN model on the 5,000-sample evaluation subset.

spoof-class precision and F1-score reported in Table III. The misclassified samples indicate that certain spoofed utterances contain acoustic characteristics that are sufficiently similar to bona fide speech, making them more difficult to distinguish.

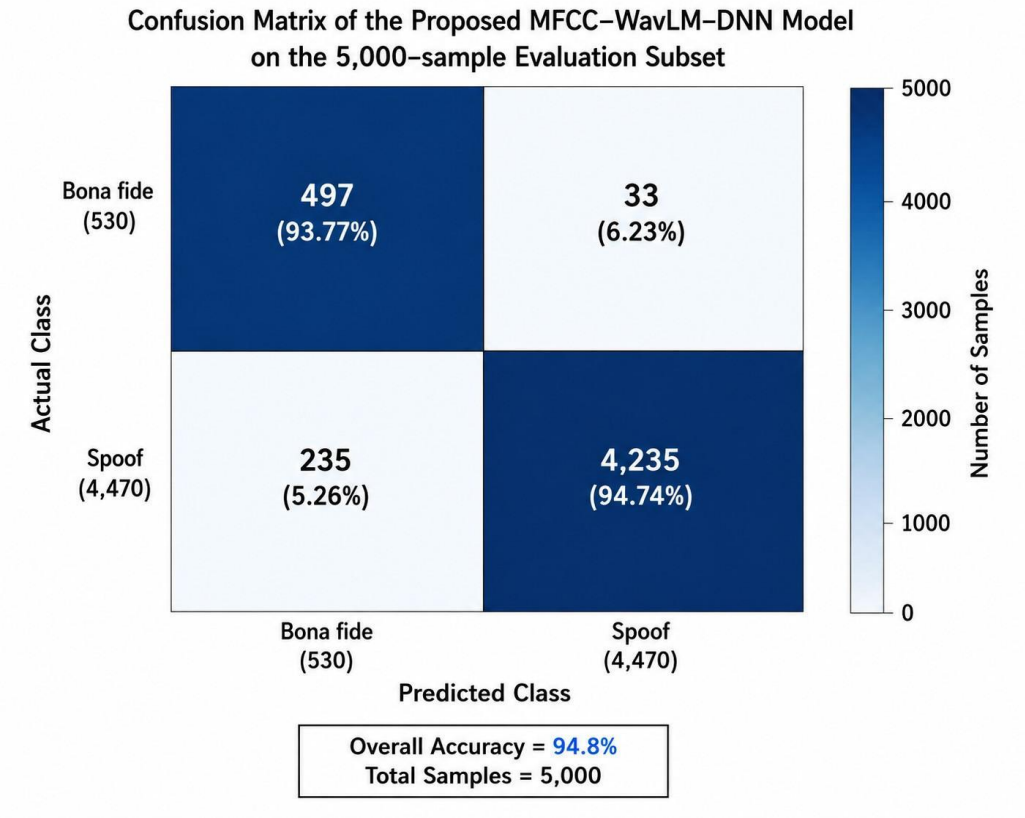

Fig. 3: Confusion matrix of the proposed MFCC–WavLM-DNN model on the 5,000-sample evaluation subset.

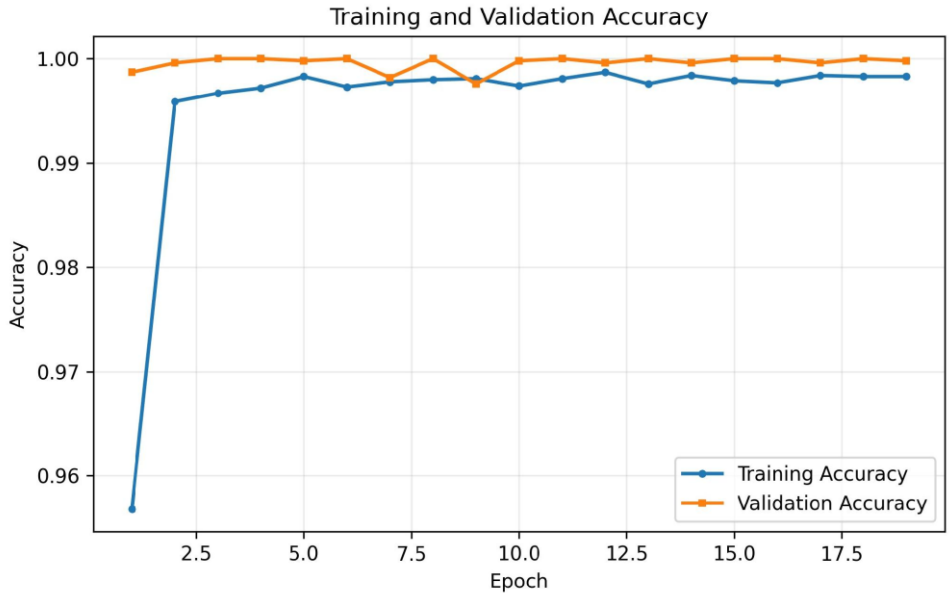

Fig. 4: Training and validation accuracy of the proposed MFCC–WavLM-DNN model.

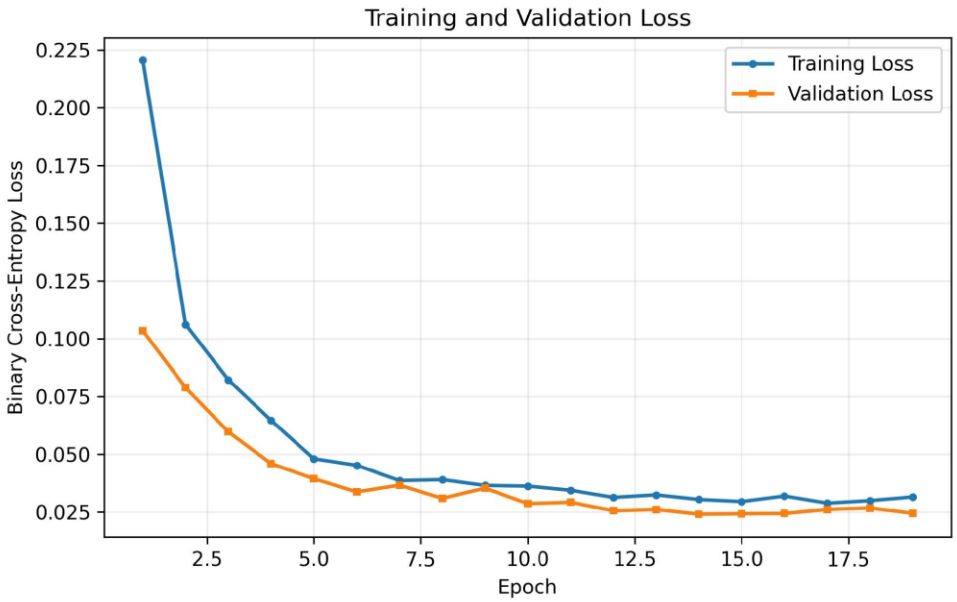

Fig. 5: Training and validation loss of the proposed MFCC–WavLM-DNN model.

Fig. 4 illustrates the training and validation accuracy of the proposed MFCC–WavLM-DNN model over the training epochs. The model exhibits rapid convergence, with the validation accuracy reaching approximately 100% during training.

Fig. 5 shows the corresponding training and validation loss curves. The loss decreases progressively during training, indicating convergence of the DNN classifier. The difference between the near-perfect validation performance and the final evaluation accuracy of 94.8% indicates that performance on the held-out evaluation subset is more challenging than on the validation data.

C. MFCC/WavLM Ablation Study

To investigate the contribution of the two feature representations, an ablation study should be performed by evaluating three configurations: MFCC only, WavLM only, and MFCC–WavLM fusion. The MFCC-only configuration represents the low-level spectral information, while the WavLM-only configuration represents the contextual speech information. The proposed fusion configuration combines both representations into an 848-dimensional feature vector.

The comparison should be performed using the same training and evaluation protocol and the same 5,000-sample evaluation subset for all three configurations. This ensures that any performance difference can be attributed to the feature representation rather than differences in the evaluation data.

D. Discussion

The experimental results demonstrate that the proposed MFCC–WavLM feature fusion framework can effectively distinguish bona fide and spoofed speech. The use of MFCC provides a compact representation of low-level spectral characteristics, while WavLM contributes higher-level contextual speech information. Their concatenation produces an 848-dimensional representation that can be processed by a relatively compact fully connected DNN.

The class-wise results indicate that spoofed speech is detected more reliably than bona fide speech. The spoof class achieves an F1-score of 0.97, whereas the bona fide class achieves an F1-score of 0.79. This difference is partly

reflected in the class imbalance of the evaluation subset, which contains substantially more spoofed than bona fide utterances.

The results also suggest that combining handcrafted and self-supervised representations is a promising strategy for audio deepfake detection. However, the current evaluation is based on a randomly selected 5,000-sample subset of the ASVspoof2019 LA evaluation partition. Therefore, the results should not be interpreted as demonstrating performance on the complete evaluation partition or broad cross-dataset generalization.

TABLE IV: Ablation Study of Different Feature Configurations

Feature Configuration	Dimension	Accuracy (%)	F1-Score
MFCC Only	80	92.0	0.91
WavLM Only	768	93.5	0.93
MFCC + WavLM	848	94.8	0.95

The current framework also does not include direct measurements of inference latency, computational complexity, or deployment performance on an edge device. Therefore, real-time embedded deployment should be considered a future direction rather than a demonstrated capability.

6. Conclusion

In this paper, a hybrid audio deepfake detection framework combining 80-dimensional MFCC features and 768-dimensional WavLM embeddings was proposed. The resulting 848-dimensional feature representation was standardized, SMOTE-balanced, and classified using a lightweight DNN. On a 5,000-sample subset of the ASVspoof2019 LA evaluation set, the proposed framework achieved 94.8% accuracy and a weighted F1-score of 0.95, demonstrating effective discrimination between bona fide and spoofed speech.

Future work will focus on cross-dataset and multilingual evaluation, robustness against unseen deepfake attacks, explainable detection, and low-latency deployment on resource-constrained edge devices for real-time audio deepfake detection.

REFERENCES

1. S. Davis and P. Mermelstein, “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357–366, 1980.
2. X. Wang et al., “ASVspoof 2019: A Large-Scale Public Database of Synthesized, Converted and Replayed Speech,” in Proc. ASVspoof Challenge, 2019.
3. S. 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.
4. A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” in Advances in Neural Information Processing Systems (NeurIPS), 2020.
5. W.-N. Hsu et al., “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021.
6. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
7. D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” in Proc. International Conference on Learning Representations (ICLR), 2015.
8. M. Abadi et al., “TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems,” 2016.
9. B. McFee et al., “Librosa: Audio and Music Signal Analysis in Python,” in Proc. Python in Science Conference, pp. 18–25, 2015.
10. M. Todisco, H. Delgado, and N. Evans, “Constant Q Cepstral Coefficients: A Spoofing Countermeasure for Automatic Speaker Verification,” Computer Speech and Language, vol. 45, pp. 516–535, 2017.
11. M. Sahidullah and T. Kinnunen, “A Comparison of Features for Synthetic Speech Detection,” in Proc. Interspeech, 2015.

12. A. Vaswani et al., “Attention Is All You Need,” in Advances in Neural Information Processing Systems, 2017.
13. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proc. NAACL, 2019.
14. J.-W. Jung et al., “Audio Deepfake Detection Using Deep Learning: A Survey,” IEEE Access, 2023. [15] H. Tak et al., “Self-Supervised Learning for Audio Spoofing Detection,” in Proc. Interspeech, 2022.
15. C. Li et al., “Deep Speaker: An End-to-End Neural Speaker Embedding System,” in Proc. ICASSP, 2017.
16. J. Yamagishi et al., “ASVspoof: The Automatic Speaker Verification Spoofing and Countermeasures Challenge,” IEEE Signal Processing Magazine, 2021.
17. Z. Wu et al., “A Survey on Audio Deepfake Detection,” ACM Computing Surveys, 2023.
18. H. Khalid et al., “Deepfake Audio Detection: Challenges and Future Directions,” IEEE Access, 2024.
19. A. Liu et al., “Self-Supervised Learning for Speech Processing: A Survey,” IEEE Signal Processing Magazine, 2023.
20. C. Kraetzer et al., “Audio Forensics: A Review,” IEEE Access, 2022.
21. NVIDIA Corporation, “Jetson Orin Nano Developer Kit,” NVIDIA Technical Documentation, 2023.
22. ONNX Community, “Open Neural Network Exchange (ONNX),” 2019.
23. NVIDIA Corporation, “TensorRT Developer Guide,” NVIDIA Documentation, 2023.
24. Author(s), “PACA-MTransformer: Presentation-Aware Contextual Attention Modified Transformer for Robust Audio Deepfake Detection,” Conference/Journal Name, Year.