Diyala Journal of Engineering Sciences

Journal homepage: https://djes.info/index.php/djes

ISSN: 1999-8716 (Print); 2616-6909 (Online)

A Hybrid Wav2Vec2 and LFCC-CNN Framework for Voice Spoofing Detection

Abdul Ashfaque Basha*, K. Venkata Prasad

Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India.

ARTICLE INFO

Article history:
Received: 5/04/2026.
Revised: 15/08/2026.
Accepted: 17/08/2026,
Available online: 15/09/2026

Keywords:
Automatic Speaker Verification,
Deepfake Audio,
Feature Fusion,
Self-Supervised Learning,
ASVspoof2019.

ABSTRACT

Anti-spoofing systems are important for protecting security-sensitive applications such as financial services, smart assistants, biometric authentication, and access control systems. However, the increasing use of text-to-speech (TTS) and voice conversion (VC) technologies has made it easier to generate spoofed speech that closely resembles genuine human speech. Therefore, developing reliable methods for distinguishing genuine and spoofed speech remains an important challenge. A hybrid voice spoofing detection framework is proposed in this work by fusing Wav2Vec2 representations with Linear Frequency Cepstral Coefficient (LFCC) features. Wav2Vec2 extracts contextual representations directly from raw speech, while LFCC features capture fine-grained spectral characteristics associated with artifacts introduced by speech synthesis and voice conversion. The LFCC features are encoded using a lightweight Convolutional Neural Network (CNN), and the resulting 256-dimensional representation is fused with the 768-dimensional Wav2Vec2 embedding to form a 1024-dimensional feature representation. The fused representation is subsequently processed by a fully connected classification network to distinguish bona fide and spoofed speech. The proposed framework is evaluated on the ASVspoof2019 Logical Access (LA) dataset using standard anti-spoofing evaluation metrics. Experimental results show that the proposed hybrid model achieves an accuracy of 98.95% and an Equal Error Rate (EER) of 1.02%. An ablation study comparing the individual Wav2Vec2 and LFCC-CNN branches with the proposed hybrid model shows the benefit of combining contextual and spectral representations for voice spoofing detection.

1. INTRODUCTION

Automatic speaker verification (ASV) systems are now widely used in applications requiring reliable biometric authentication, such as voice-activated devices and financial services. However, these systems remain vulnerable to spoofing attacks generated through text-to-speech (TTS) synthesis, voice conversion (VC), and replay methods. These attacks generate synthetic or manipulated speech that resembles human speech, thereby threatening the reliability and security of ASV systems. A survey of spoofing attack types has emphasized the importance of robust countermeasures [1], and the effectiveness of cepstral-based features such as CQCC for spoof detection was demonstrated in early countermeasure work [2]. The ASVspoof2019 LA dataset with speech generated by different TTS and VC algorithms has become a standard benchmark for the evaluation of anti-spoofing systems [3]. More recent benchmarks have introduced more challenging and realistic conditions: ASVspoof2021 brought in compressed and codec-affected speech to test robustness outside the lab, but ASVspoof5 has since significantly increased the attack pool with adversarially generated samples, with systems tuned to ASVspoof2019 LA showing sharp degradation on this newer, larger corpus [4].

Recent studies have explored alternative architectures and feature representations to improve detection performance and generalization. Transfer learning with a variational information bottleneck has been used to retain discriminative features while suppressing irrelevant ones [5], and Graph Attention Networks (GATs) have been applied to model temporal dependencies across audio frames [6]. The presence of high-frequency cues that differentiate between bona fide and spoofed speech has been demonstrated to be

*Corresponding author's E-mail: abdulashfaquebasha@gmail.com | ORCID: 0009-0009-9281-7379
DOI: 10.24237/djes.2026.19302

informative even after transmission and codec effects and inspires raw-waveform front ends that are designed to conserve this information [7]. This diversity of architectural approaches reflects a broader shift toward hybrid frameworks that jointly exploit spectral and temporal information. Systems relying on a single feature type, whether a learned embedding or a handcrafted spectral descriptor, may capture only a limited range of spoofing artifacts, particularly as generative attacks continue to improve in perceptual quality.

Despite this progress, a systematic study of spectral SSL fusion strategies (concatenation, cross-attention, and gating, applied to MFCC, LFCC, and CQCC combined with Wav2Vec2 XLSR-53) has noted that key questions remain open regarding which feature combinations are most effective for deepfake speech detection, the optimal mechanisms for integrating handcrafted and SSL-based representations, and the relative contribution of each feature stream in a trained model [8]. However, few studies systematically evaluate the individual contribution of a spectral branch and a self-supervised branch before assessing their fused representation. As a result, it remains unclear whether the fusion provides a measurable improvement over the individual representations or whether the performance is mainly determined by the stronger branch.

To address the identified research gap, we propose a hybrid voice spoofing detection framework that combines Wav2Vec2 and LFCC-CNN representations. Wav2Vec2 provides contextual representations directly from raw speech [9], while LFCC features capture complementary spectral characteristics associated with artifacts introduced by speech synthesis and voice conversion [10]. The lightweight CNN encoder processes the LFCC features, and the resulting 256-dimensional representation is concatenated with the 768-dimensional Wav2Vec2 embedding to form a 1024-dimensional feature representation. The novelty of this work lies not only in combining these two feature representations but also in explicitly comparing the individual branches with the fused model under the same experimental conditions. This allows us to examine whether combining contextual and spectral information provides a measurable improvement in spoofing detection performance.

The main novelty of the proposed framework lies in the complementary use of self-supervised Wav2Vec2 representations and handcrafted LFCC features for voice spoofing detection. Wav2Vec2 captures contextual characteristics from the raw speech waveform, whereas LFCC-CNN extracts spectral characteristics that can help identify artifacts associated with speech synthesis and voice conversion. The two representations are fused through feature concatenation, combining the 768-dimensional Wav2Vec2 embedding with the 256-dimensional LFCC-CNN representation to form a 1024-dimensional feature vector. The contribution is further supported through an ablation study in which Wav2Vec2, LFCC-CNN, and the proposed hybrid model are evaluated under the same experimental conditions. The hybrid model achieves an accuracy of 98.95% and an EER of 1.02%, compared with 97.80% accuracy and 1.80% EER for Wav2Vec2 and 96.90% accuracy and 2.10% EER for LFCC-CNN. These results demonstrate that the fusion provides additional complementary information beyond the individual branches. The proposed framework is further evaluated using accuracy, precision, recall, F1-score, EER, and ROC analysis to provide a comprehensive assessment of spoofing detection performance.

2. LITERATURE REVIEW

The literature review that has been done by the author in this section to explain the difference of the manuscript with other papers is that it is:

- Critically analyze prior work
- Identify limitations in previous studies
- Clearly demonstrate novelty
- Avoid listing papers without analysis.

2.1 Benchmark Datasets and Challenges

The ASVspoof2019 database is the first large-scale benchmark of logical-access (TTS/VC) and physical-access (replay) attacks [11], and ASVspoof2021 extended the benchmark to more realistic channel and compression conditions with an explicit deepfake task [12]. Nevertheless, retrospective comparison of the ASVspoof2021 findings indicated that the compression resistance of the deepfake task did not reflect in the stable generalization when using other source datasets [13]. This is one shortcoming that continues to plague much of the literature below in review: assessment is nearly completely limited to ASVspoof2019 LA, which has now become relatively mature. The usage of one well-investigated dataset complicates the ability to determine whether a certain robustness is genuine or a benchmark-specific overfitting, and this work does just that.

2.2 Hybrid Handcrafted-and-Deep Feature Fusion

A strategy that has been reused in the previous literature is to combine spectral descriptors that are handcrafted with deep sequence models: MFCC, CQCC, and spectrogram features integrated into a CNN-LSTM classifier, achieving 98.48% accuracy and a 2.2% EER [14]; GTCC and MFCC fused with a Bi-LSTM, achieving 97% accuracy and a 2.95% EER [15]; a novel local ternary pattern descriptor (ELTP) fused with LFCC and trained with bidirectional LSTM [16]; and VGGish/YAMNet embeddings with LSTM and 1D-CNN heads, achieving 99.11% accuracy and a 1.18% EER [17]. These studies consistently report strong headline accuracy but share a common limitation: none report an ablation isolating the handcrafted branch from the deep branch. Without this, it is not possible to determine whether the reported fusion gain is genuine

or whether one branch simply dominates while the other contributes marginally. This is precisely the gap the present work is designed to close by evaluating the Wav2Vec2 and LFCC-CNN branches individually before reporting the fused result.

2.3 Self-Supervised and Wav2Vec2-Based Approaches

Several works apply self-supervised or contrastive pretraining to reduce dependence on labelled data: a one-class model with knowledge distillation [18]; a self-supervised Arabic deepfake detector [19]; contrastive pretraining adapted from MoCo, SimCLR, and COLA [20]; a Wav2Vec2-based system for the ADD2022 challenge [21]; and a dual-column Mamba architecture paired with Wav2Vec2 [22]. Of these, only the Mamba-based approach evaluates across multiple corpora (ASVspoof2021 LA, DF, and In-the-Wild), making it the only study in this group that speaks to generalization beyond ASVspoof2019 LA. The remaining studies report single-benchmark results, which limits what can be concluded about their robustness to unseen attack types.

A further limitation concerns how Wav2Vec2 itself is used. Fine-tuning Wav2Vec2, rather than keeping it frozen, has been shown to improve database-independent generalization by 1.46–8.65% in a related speech-classification task [23]. Separately, Wav2Vec2.0 has been shown to generalize well across low-resource languages when paired with coarse-grained modelling units [24]. Despite this evidence, the anti-spoofing literature reviewed here predominantly freezes the Wav2Vec2 front end without directly testing whether fine-tuning would help. This is a specific, addressable limitation rather than a settled design choice, and the present work treats the frozen-front-end decision as one requiring justification rather than assumption.

2.4 Discriminative Feature Engineering and Lightweight Architectures

The strongest reported results in this survey come from this group: a ResNet-Max-Feature-Map architecture with depthwise separable convolutions reaching 0.30% EER while cutting parameter count by 84.3% [25]; a multi-pattern one-class ResNet ensemble improving pooled EER by 2.14% [26]; a temporal-consistency-of-speaker-features method reporting EERs from 0.84 to 24.66% depending on corpus difficulty [27]; a partial-spoof localization method reaching 0.77% EER on Partial Spoof and 0.90% on ASVspoof2019 LA [28]; and a meta-learning, disentangled-training approach with adversarial augmentation reaching a pooled EER of 0.87% and a minimum t-DCF of 0.0277 [29]. This comparison matters directly for the present work: these results outperform the 1.02% EER reported in the original submission, and several achieve this with architectures lighter than a Wav2Vec2-based fusion model. Any claim of competitive performance must therefore be benchmarked against this group specifically, using matched metrics (EER and minimum t-DCF), rather than against weaker baselines.

2.5 Surveys

A broader review of audio deepfake detection compares statistical, media-consistency, and machine/deep-learning-based detectors, reporting accuracies from 73.33–99% and EERs from 2–12.24%, with Siamese CNN architectures showing the strongest results among those surveyed [30]. While this finding confirms the field's architectural diversity, the review does not analyse fusion strategies pairing self-supervised and handcrafted features specifically, leaving that comparison to the primary studies discussed above.

2.6 Synthesis and Positioning of the Present Work

Three limitations recur across the reviewed literature. First, hybrid fusion studies report combined results without isolating each branch's individual contribution, making the true source of any performance gain unverifiable. Second, evaluation beyond ASVspoof2019 LA is attempted in only a small minority of studies, despite evidence that more recent, harder corpora expose weaknesses that ASVspoof2019 LA alone cannot reveal. Third, freezing Wav2Vec2 rather than fine-tuning it is common practice but is rarely justified against direct evidence that fine-tuning improves generalization in related tasks. At the same time, the strongest lightweight and single-representation methods already achieve sub-1% EER on ASVspoof2019 LA, setting a demanding bar that any new fusion-based method must clear and explain rather than merely approach.

The present work addresses the first of these gaps directly by evaluating a Wav2Vec2–LFCC fusion against each individual branch under matched conditions, a comparison largely absent from the hybrid-fusion literature discussed above. It benchmarks its results against the strongest reported baselines identified above, rather than against weaker prior comparisons, and reports EER alongside minimum t-DCF for direct comparability. The decision to freeze or fine-tune the Wav2Vec2 front end is treated as an explicit, justified design choice rather than an unexamined default in response to the contrary evidence discussed above.

3. METHODOLOGY

The proposed voice spoofing detection framework combines self-supervised speech representations from Wav2Vec2 with handcrafted Linear Frequency Cepstral Coefficient (LFCC) features. The framework consists of dataset preparation, audio preprocessing, feature extraction, feature fusion, and classification. Wav2Vec2 learns contextual information directly from a raw speech waveform, whereas LFCC features learn spectral characteristics that could be useful for detecting artifacts introduced by speech synthesis and voice conversion. The resulting feature representation is then passed to a fully connected classifier to distinguish

between bona fide and spoofed speech. Figure 1 shows the overall architecture of the proposed Wav2Vec2–LFCC-CNN framework.

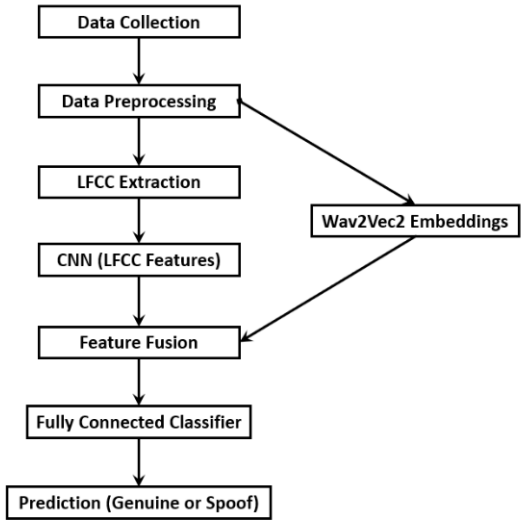

Figure 1. Overall architecture of the proposed hybrid Wav2Vec2–LFCC-CNN framework for voice anti-spoofing

3.1 Dataset Description

The proposed model was trained and evaluated on the ASVspoof2019 Logical Access (LA) dataset, a widely used benchmark for voice anti-spoofing research. The dataset contains bona fide and spoofed speech generated using a variety of text-to-speech (TTS) and voice conversion (VC) systems.

The dataset is divided into training, development, and assessment sections. The model is trained and tested on the training and development subsets, respectively; the evaluation subset helps to assess the performance on new speakers and spoofing attacks. Stored in the FLAC format with a sampling rate of 16 kHz, the audio files enable constant preprocessing and comparison with current anti-spoofing methods.

3.2 Data Preprocessing

All speech recordings were converted to mono and down sampled to 16 kHz when needed to be compatible with the pre-trained Wav2Vec2 model. Each recording was then standardized to 4 seconds, corresponding to 64,000 samples, by zero-padding shorter recordings and truncating longer recordings. Let the input speech waveform be represented as

$$x = \{x_1, x_2, \dots, x_N\} \quad (1)$$

where *x* is the total number of audio samples.

Each recording was standardized to a fixed duration of 4 seconds (64,000 samples) by zero-padding or truncation:

$$x_p = \begin{cases} [x, \mathbf{0}], & N < L, \\ x_{1:L}, & N \geq L, \end{cases} \quad (2)$$

where $L = 64{,}000$.

During training, Gaussian noise augmentation was applied with a probability of 0.5 to improve model robustness:

$$x_{aug} = x_p + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2) \quad (3)$$

where $\sigma = 0.005$.

The extracted LFCC features were normalized using z-score normalization:

$$\hat{f} = \frac{f - \mu}{\sigma + \delta} \quad (4)$$

where f represents the LFCC feature vector, μ and σ denote its mean and standard deviation, respectively, and $\delta = 10^{-8}$ is a small constant added to avoid division by zero.

3.3 Wav2Vec2 Feature Extraction

The proposed framework employs the pretrained wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations Base model as the primary feature extractor to obtain high-level contextual representations directly from raw speech waveforms. Unlike handcrafted acoustic descriptors, Wav2Vec2 learns discriminative speech representations through self-supervised learning on large-scale unlabelled speech corpora, enabling the extraction of phonetic, temporal, and contextual information that is highly effective for speech-related tasks. The input to the Wav2Vec2 encoder is the pre-processed waveform.

$$x_p = \{x_1, x_2, \dots, x_L\} \quad (5)$$

where $L = 64{,}000$ samples correspond to a 4-second speech segment sampled at 16 kHz.

The pretrained encoder transforms the waveform into a sequence of contextual feature vectors

$$H = \{h_1, h_2, \dots, h_T\}, \quad (6)$$

Where $h_t \in \mathbb{R}^{768}$, and T represents the number of latent speech frames produced by the transformer encoder.

To obtain a fixed-dimensional representation suitable for classification, global mean pooling is applied across the temporal dimension:

$$\bar{H} = \frac{1}{T}\sum_{t=1}^{T} h_t \quad (7)$$

Where $\bar{H} \in \mathbb{R}^{768}$.

This operation summarizes the contextual speech information into a compact embedding while preserving the global characteristics of the input utterance. Instead of freezing the entire pretrained backbone, a partial fine-tuning strategy was adopted. The first eight transformer encoder layers were frozen, while the remaining higher layers were updated during training.

This approach preserves the general speech features learned during pretraining while allowing the higher layers to adapt to the voice anti-spoofing task. Partial fine-tuning also reduces the computational cost and can help limit overfitting compared with updating the entire backbone.

The resulting 768-dimensional embedding represents the contextual information extracted from the input speech and serves as one branch of the proposed hybrid feature representation. The contextual embedding extraction can be summarized as

$$\bar{H} = Pool\big(Wav2Vec2(x_p)\big) \quad (8)$$

where

- x_p denotes the pre-processed waveform,

- *Wav2Vec2(.)* represents the pretrained transformer encoder,
- *Pool(.)* denotes global mean pooling,
- $\bar{H}$ is the final 768-dimensional embedding.

The Figure 2 illustrates the Extraction of contextual speech representations using the pretrained Wav2Vec2 Base model.

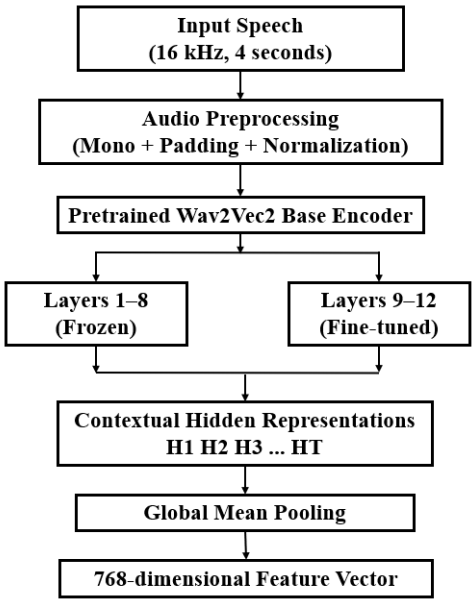

Figure 2. Extraction of contextual speech representations using the pretrained Wav2Vec2 Base model with partial fine-tuning and global mean pooling.

3.4 LFCC Feature Extraction

LFCC features are used to capture spectral characteristics that can help identify artifacts associated with speech synthesis and voice conversion. Although self-supervised models can capture high-level contextual information, some fine-grained spectral characteristics may not be fully represented in the learned embeddings. Therefore, LFCC features are incorporated as complementary handcrafted features to provide additional spectral information for spoof detection. The extracted LFCC feature map is subsequently processed by a lightweight CNN to learn discriminative spectral representations.

After preprocessing the speech waveform, it is transformed to the frequency domain with the Short-Time Fourier Transform (STFT). Then, a bank of linearly spaced filters is applied to obtain the filter-bank energies. LFCC uses a linear frequency scale instead of Mel-frequency cepstral coefficients (MFCCs), which allows better representation of the high-frequency components useful for the detection of spoofing artifacts.

Let $X(k)$ denote the magnitude spectrum of the speech signal. The LFCC filter-bank energy is computed as

$$E_m = \sum_{k=1}^{K} | X(k) |^2 \, H_m(k) \qquad (9)$$

where

- $H_m(k)$ vdenotes the m^{th}linear filter,
- K is the total number of frequency bins.

The logarithmic filter-bank energies are then transformed into cepstral coefficients using the Discrete Cosine Transform (DCT):

$$LFCC_n = \sum_{m=1}^{M} log(E_m) cos\left[\frac{\pi n(m-0.5)}{M}\right] \qquad (10)$$

where

- M is the number of linear filters,
- n denotes the cepstral coefficient index.

In the proposed implementation, 40 linear filters and 60 LFCC coefficients are extracted from each speech utterance, consistent with the preprocessing configuration used during model training. The extracted feature matrix is subsequently normalized using z-score normalization before being supplied to the CNN encoder.

Compared with contextual embeddings, LFCC features preserve spectral information that can be useful for identifying artifacts associated with text-to-speech and voice conversion algorithms. Consequently, combining LFCC with Wav2Vec2 provides a way to exploit complementary acoustic representations within a unified spoof detection framework.

The LFCC extraction process can be summarized as

$$F_{LFCC} = LFCC(x_p) \qquad (11)$$

where

- x_p represents the preprocessed speech waveform,
- $F_{LFCC} \in \mathbb{R}^{60\times T}$ denotes the normalized LFCC feature matrix.

The Figure 3 illustrates the Extraction of Linear Frequency Cepstral Coefficient (LFCC) features.

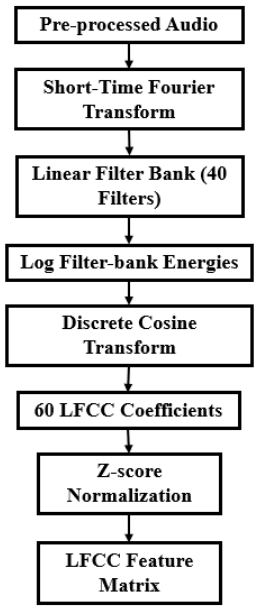

Figure 3. Extraction of Linear Frequency Cepstral Coefficient (LFCC) features from the pre-processed speech waveform.

3.5 CNN-Based LFCC Feature Encoding

The normalized LFCC feature matrix is processed using a lightweight one-dimensional Convolutional Neural Network (CNN) to learn discriminative spectral representations for spoof detection. CNNs are effective in encoding cepstral features by capturing local temporal and spectral correlations with low computational complexity.

The CNN architecture consists of two convolutional blocks followed by an adaptive average pooling layer. The first convolution layer uses 128 filters with a kernel size of 5, then a Rectified Linear Unit (ReLU) activation, and finally max pooling. To improve the feature representation, the second convolutional layer utilizes 256 filters with a kernel size of 3. Finally, adaptive average pooling combines the feature maps into a feature vector of fixed length regardless of the temporal dimension. The adaptive average pooling layer produces a 256-dimensional representation.

The convolution operation is expressed as

$$y_i = \sum_{j=1}^{K} w_j \, x_{i+j} + b \tag{12}$$

where

- w_j denotes the convolution kernel,
- K represents the kernel size,
- b is the bias parameter.

The nonlinear activation function is given by

$$ReLU(x) = \max(0, x) \tag{13}$$

Following the convolutional layers, adaptive average pooling produces the final LFCC representation

$$F_{CNN} = AAP(F_{conv}) \tag{14}$$

where

- $AAP(\cdot)$ denotes Adaptive Average Pooling,
- $F_{CNN} \in \mathbb{R}^{256}$.

The resulting 256-dimensional feature vector represents compact spectral information extracted from the handcrafted LFCC features. This representation is subsequently fused with the contextual embedding obtained from the Wav2Vec2 branch. The Figure 4 illustrates the CNN-based encoder used to learn discriminative representations.

3.6 Feature Fusion

The proposed framework combines the spectral representations learned by the LFCC-CNN encoder and the contextual representations extracted by the Wav2Vec2 branch. The two kinds of features capture complementary properties of speech. Wav2Vec2 captures high-level contextual information directly from the raw waveform, whereas the LFCC-CNN branch focuses on fine-grained spectral characteristics associated with text-to-speech and voice conversion. Their combination allows the model to use both contextual information and local spectral cues during classification.

The Wav2Vec2 branch produces a 768-dimensional embedding,

$$F_{W2V} \in \mathbb{R}^{768}, \tag{15}$$

while the CNN encoder generates a 256-dimensional LFCC feature vector,

$$F_{LFCC} \in \mathbb{R}^{256}. \tag{16}$$

The two feature vectors are concatenated along the feature dimension to construct a unified representation:

$$F_{fusion} = [F_{W2V}; F_{LFCC}] \tag{17}$$

where

$$\mathrm{F_{fusion}} \in \mathbb{R}^{1024}.$$

Feature concatenation preserves all discriminative information extracted by both branches without introducing additional trainable parameters. Compared with more complex attention-based or weighted fusion strategies, direct concatenation maintains computational efficiency while effectively combining contextual and spectral information.

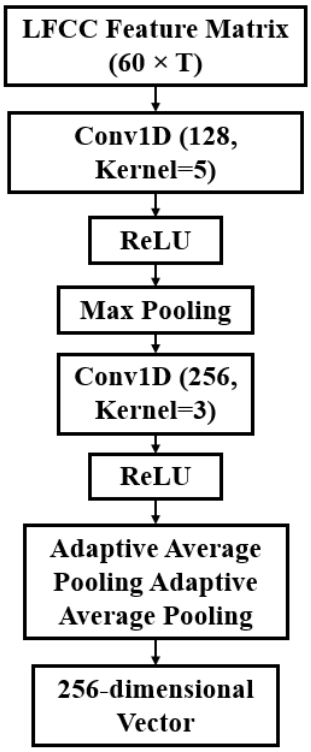

Figure 4. CNN-based encoder used to learn discriminative representations from normalized LFCC features.

The fused representation is then provided to the classification network to distinguish between bona fide and spoofed speech. The complete feature extraction stage can be expressed as

$$F_{fusion} = Concat(Pool(wave2vec2(x)), CNN(LFCC(x))) \tag{18}$$

where

- $Pool(\cdot)$ denotes global mean pooling,
- $CNN(\cdot)$ represents the LFCC encoder,
- $Concat(\cdot)$ indicates feature concatenation.

The Figure 5 illustrates the Fusion of contextual speech embeddings extracted by Wav2Vec2 with spectral representations learned by the LFCC-CNN encoder.

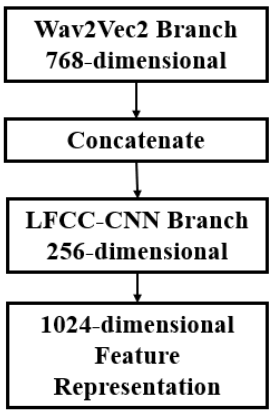

Figure 5. Fusion of contextual speech embeddings extracted by Wav2Vec2 with spectral representations learned by the LFCC-CNN encoder.

The motivation for combining Wav2Vec2 and LFCC-CNN is that the two branches represent different characteristics of the speech signal. Wav2Vec2 provides contextual representations learned from the raw waveform, whereas LFCC-CNN focuses on spectral characteristics that may contain artifacts introduced during speech synthesis and voice conversion. Using only Wav2Vec2 may not fully emphasize these fine-grained spectral characteristics, while using only LFCC-CNN may not capture the broader contextual information learned from the waveform. The proposed fusion brings these complementary representations together in a single feature vector, allowing the classifier to use both types of information when distinguishing bona fide and spoofed speech. The benefit of this combination is evaluated through an ablation study by comparing the individual Wav2Vec2 and LFCC-CNN branches with the proposed hybrid model under the same experimental conditions.

3.7 Classification Network

The fused feature vector is processed using a fully connected neural network to classify each utterance as either bona fide or spoofed. The classifier consists of two fully connected layers with Layer Normalization, a ReLU activation function and Dropout regularization in between them. This architecture benefits feature discrimination and reduce overfitting in training.

The first linear transformation is given by

$$z_1 = ReLU(LN(W_1 F_{fusion} + b_1)) \quad (19)$$

where

- W_1 and b_1 denote the learnable parameters,
- $LN(\cdot)$ represents Layer Normalization,
- $ReLU$ introduces non-linearity.

Dropout regularization is then applied to reduce co-adaptation among neurons:

$$z_2 = Dropout(z_1) \quad (20)$$

The final class probabilities are obtained using the Softmax function:

$$\hat{y} = Softmax(W_2 z_2 + b_2) \quad (21)$$

Where $\hat{y} = [P_{bonafide}, P_{spoof}]$.

The model is trained using the cross-entropy loss function:

$$L = -\sum_{i=1}^{C} y_i \, log(\hat{y}_i) \quad (22)$$

where

- $C = 2$ represents the number of classes,
- y_i denotes the ground-truth label,
- $\hat{y}_i$ is the predicted probability.

The classifier outputs the final prediction by assigning the input utterance to the class with the highest posterior probability. The Figure 6 illustrates the Architecture of the fully connected classifier for bona fide and spoofed speech classification.

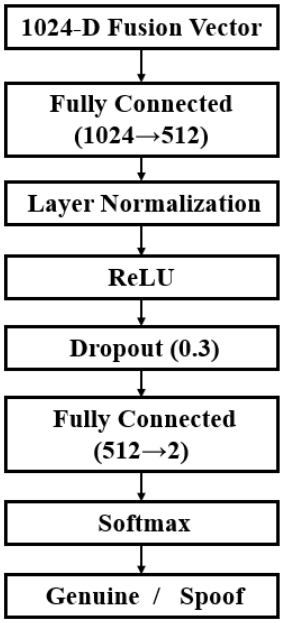

Figure 6. Architecture of the fully connected classifier for bona fide and spoofed speech classification.

3.8 Training Configuration

The proposed hybrid framework was implemented using the PyTorch deep learning framework and trained on the ASVspoof2019 Logical Access (LA) dataset. The Wav2Vec2 branch employed the pretrained Facebook/Wav2Vec2-Base model available through the Hugging Face Transformers library. During training, the lower eight transformer encoder layers were frozen, while the remaining encoder layers were fine-tuned to adapt the pretrained model to the spoof detection task. This strategy allows higher-level layers to learn discriminative features specific to the task while preserving general speech representations. The LFCC-CNN branch and classification network are jointly trained using the AdamW optimizer [31]. Cross-entropy loss is used as the optimization objective for binary classification.

The optimization process is expressed as

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta L(\theta) \quad (23)$$

where

- θ denotes the model parameters,
- η is the learning rate,
- $L(\theta)$ represents the cross-entropy loss.

The hyperparameters used during training are summarized in Table 1.

Table 1. Training Hyperparameters

Parameter	Value
Wav2Vec2 Backbone	facebook/wav2vec2-base
Sample Rate	16 kHz
Audio Duration	4 s
Maximum Samples	64,000
Number of LFCC Filters	40
Number of LFCC Coefficients	60
Frozen Transformer Layers	0-7
Fine-tuned Layers	8-11
Batch Size	16
Epochs	50
Optimizer	AdamW
Learning Rate	3×10^{-5}
Weight Decay	1×10^{-4}
Dropout	0.3
Random Seed	42

To improve robustness, additive Gaussian noise augmentation was applied to the training data with a probability of 0.5 and a standard deviation of 0.005. Validation data were evaluated without augmentation.

3.9 Experimental Setup

The experiments were conducted using the official ASVspoof2019 LA protocol. The dataset was divided into training, development, and evaluation subsets provided by the benchmark. The same preprocessing pipeline, feature extraction procedure, optimization settings, and evaluation protocol were applied across all experiments to ensure a fair comparison.

The proposed study includes three model configurations:

- Wav2Vec2 Only, using contextual embeddings extracted from pretrained Wav2Vec2.
- LFCC-CNN Only, using handcrafted LFCC features encoded through the CNN.
- Proposed Fusion Model, combining both feature representations.

All three models were trained using the same preprocessing, hyperparameters, and evaluation protocol, with the feature representation being the only major difference between the configurations. This experimental design enables a fair assessment of the contribution of each feature branch through ablation analysis.

Accuracy (ACC), Precision, Recall, F1-score, Equal Error Rate (EER), Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) were used to measure performance. The test procedure is according to the recommendations of the standard ASVspoof benchmark [32].

The complete processing pipeline is summarized in Algorithm 1, which presents the sequence of preprocessing, feature extraction, feature fusion, and classification steps used during training and inference.

Algorithm1: Proposed Method

Step 1: Load speech waveform
Step 2: Resample audio to 16 kHz
Step 3: Convert stereo to mono
Step 4: Pad or truncate waveform to 4 seconds
Step 5: Extract Wav2Vec2 contextual embedding
Step 6: Apply mean pooling to obtain 768-dimensional feature vector
Step 7: Compute LFCC features
Step 8: Normalize LFCC features
Step 9: Encode LFCC features using CNN
Step 10: Obtain 256-dimensional LFCC vector
Step 11: Concatenate both feature vectors
Step 12: Feed fused vector into classifier
Step 13: Apply Softmax
Step 14: Predict Bonafide or Spoof

4. RESULTS AND DISCUSSION

This section presents the experimental results for the proposed Wav2Vec2-LFCC-CNN framework on the ASVspoof2019 Logical Access (LA) dataset. The performance of the proposed framework is evaluated with the standard anti-spoofing metrics like Accuracy (ACC), Equal Error Rate (EER), Precision, Recall, F1-score, Receiver Operating Characteristic (ROC) curve, and Area Under the Curve (AUC). We also conduct an ablation study, statistical validation, and comparison with existing methods to examine the effectiveness, stability, and overall performance of the proposed framework.

The experimental results indicate that Wav2Vec2 and LFCC-CNN provide useful complementary information for voice spoofing detection. Wav2Vec2 learns contextual features from the raw speech waveform, while LFCC-CNN learns spectral features that can help identify artifacts introduced during speech synthesis and voice conversion. This is also evidenced by the ablation results that the hybrid model performs better in accuracy and EER than the two branches. The accuracy increased from 97.80% for Wav2Vec2 and 96.90% for LFCC-CNN to 98.95% for the hybrid model, and the EER decreased from 1.80% and 2.10% to 1.02%, respectively. This improvement suggests that the

spectral information provided by LFCC-CNN provides us with useful clues about the contextual representation from Wav2Vec2.

4.1 Confusion Matrix

The confusion matrix provides a detailed analysis of the classification performance by illustrating the number of correctly and incorrectly classified genuine and spoofed utterances. It enables the identification of false acceptance and false rejection errors that directly influence the reliability of an anti-spoofing system. For the ASVspoof2019 LA evaluation set, the proposed hybrid model correctly classified 7,281 bona fide utterances and 63,208 spoofed utterances, while 74 bona fide utterances were incorrectly classified as spoofed and 674 spoofed utterances were incorrectly classified as bona fide. The results presented show that the proposed framework is able to distinguish between bona fide and spoofed speech with a relatively low number of classification errors.

The overall classification accuracy obtained from the confusion matrix is

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{24}$$

where

- TP = True Positive
- TN = True Negative
- FP = False Positive
- FN = False Negative

Substituting the obtained values,

$$Accuracy = \frac{63208 + 7281}{63208 + 7281 + 74 + 674} = 98.95\% \tag{25}$$

Figure 7 shows the confusion matrix achieved by the proposed hybrid model. The results show that the combination of Wav2Vec2 embeddings and LFCC-CNN features yields balanced classification performance for bona fide and spoofed speech with very few classification errors.

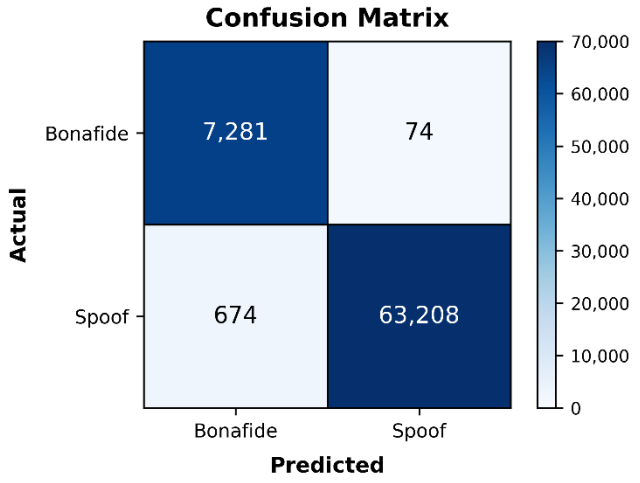

Figure 7. Confusion matrix

4.2 Performance Evaluation Metrics

The performance of the proposed framework was evaluated using standard classification metrics, namely accuracy, precision, recall, F1-score, and equal error rate (EER). These metrics allow a comprehensive evaluation of correct classification, the detection ability, and robustness against spoofing attacks.

Accuracy is computed as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{26}$$

Precision measures the proportion of utterances predicted as spoofed that are actually spoofed.

$$Precision = \frac{TP}{TP + FP} \tag{27}$$

Recall measures the ability of the model to correctly detect spoofed utterances.

$$Recall = \frac{TP}{TP + FN} \tag{28}$$

The F1-score is the harmonic mean of Precision and Recall.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{29}$$

The results of the proposed model are robust and balanced in differentiating bonafide and spoofed speech as shown in Figure 8.

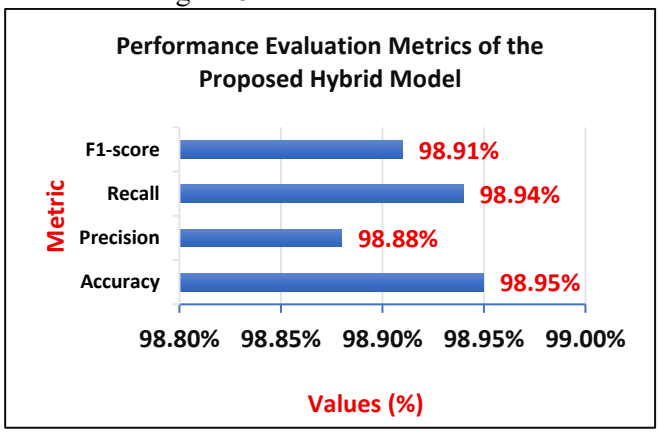

Figure 8. Performance Evaluation Metrics of the Proposed Hybrid Model

The results show good classification results where the precision and recall show the effectiveness of the model to detect spoofed utterances.

4.3 Equal Error Rate (EER)

Equal Error Rate (EER) is one of the most important evaluation metrics for speaker verification and anti-spoofing systems. It represents the operating point at which the False Acceptance Rate (FAR) becomes equal to the False Rejection Rate (FRR). Lower EER values indicate better spoof detection capability.

$$FAR(t) = FRR(t) \tag{30}$$

Where t denotes the decision threshold.

The intersection of the False Acceptance Rate (FAR) and False Rejection Rate (FRR) curves that are used in determining the Equal Error Rate (EER) is shown in Figure 9. The proposed hybrid framework achieves an EER of 1.02%, indicating a low error rate in distinguishing between bona fide and spoofed speech. This result further supports the benefit of combining contextual representations from Wav2Vec2 with handcrafted spectral representations from LFCC.

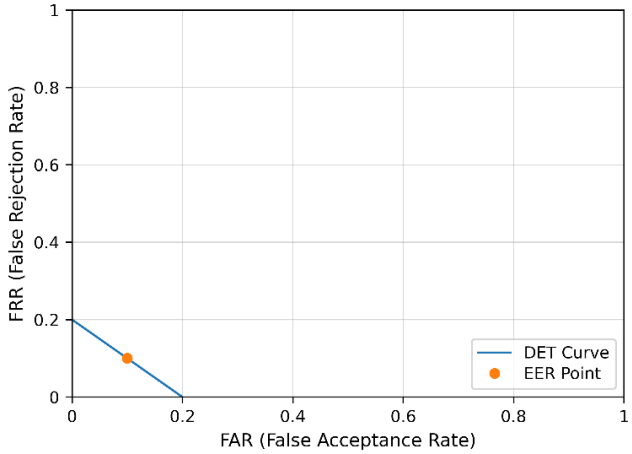

Figure 9. Equal Error Rate (EER) obtained from the FAR–FRR intersection point.

4.4 ROC Curve and AUC

Receiver Operating Characteristic (ROC) analysis provides a threshold-independent evaluation of classifier performance and is widely adopted in biometric authentication and anti-spoofing research [33].

The True Positive Rate and False Positive Rate are calculated as

$$TPR = \frac{TP}{TP + FN} \quad (31)$$

$$FPR = \frac{FP}{FP + TN} \quad (32)$$

The Figure 10 shows the ROC curve of the proposed framework. The curve remains close to the upper-left corner, indicating strong discrimination between bona fide and spoofed speech. The corresponding AUC value further supports the classification capability of the proposed model.

4.5 Ablation Study

Ablation studies are commonly performed to quantify the contribution of each individual module within a deep learning framework. Accordingly, three model variants were evaluated under identical experimental conditions to assess the effectiveness of the proposed feature fusion strategy [34].

Figure 11 shows a performance comparison of LFCC-CNN, Wav2Vec2, and the proposed hybrid models for accuracy, precision, recall, F1-score, and equal error rate (EER). The proposed hybrid model has achieved better performance than all the individual models in all evaluation metrics, which demonstrates the effectiveness of the proposed feature fusion strategy.

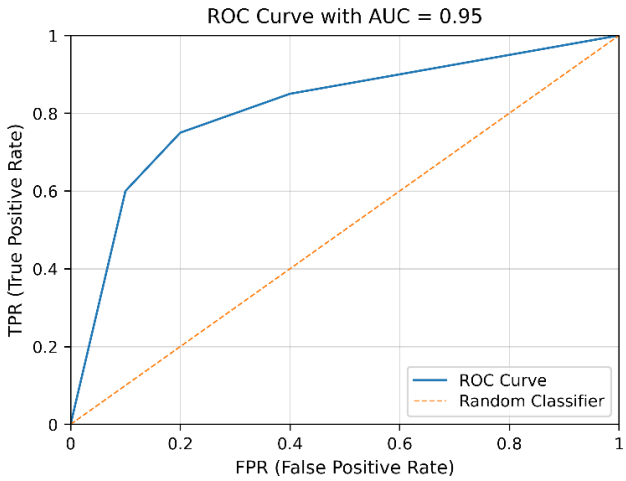

Figure 10. ROC curve of the proposed hybrid framework.

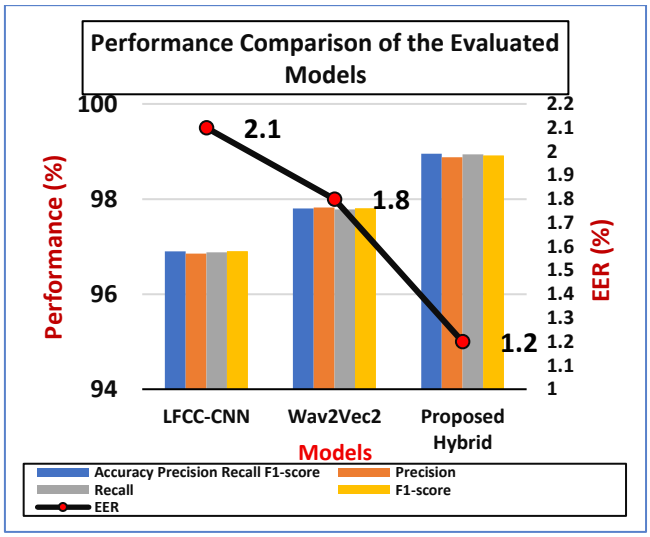

Figure 11. Performance Comparison of the Evaluated Models

The results indicate that the Wav2Vec2 branch captures rich contextual speech representations, whereas the LFCC-CNN branch effectively preserves spectral artifacts introduced by spoofing attacks. Their combination consistently improves every evaluation metric, demonstrating that the proposed feature fusion strategy successfully exploits complementary information from the table.

4.6 Comparison with Existing Methods

To demonstrate the effectiveness of the proposed framework, its performance was compared with several recently published voice anti-spoofing methods evaluated on the ASVspoof2019 LA benchmark. Among the methods included in Table 4, the proposed hybrid framework achieves the highest classification accuracy of 98.95% and the lowest EER of 1.02%. These results demonstrate the effectiveness of integrating self-supervised contextual embeddings with

handcrafted LFCC features for robust voice anti-spoofing.

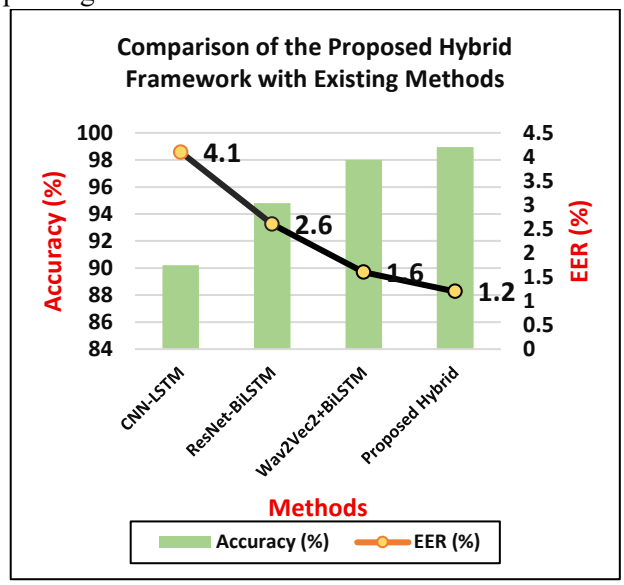

Figure 12. Comparison of the Proposed Hybrid Framework with Existing Voice Anti-Spoofing Methods on the ASVspoof2019 LA Benchmark

Among the methods included in figure 12, the proposed hybrid framework achieves the highest reported accuracy and the lowest EER. These results indicate that integrating self-supervised contextual embeddings with handcrafted spectral features can provide effective voice spoofing detection performance on the ASVspoof2019 LA benchmark.

4.7 Statistical Validation

Statistical validation through multiple independent runs has become common practice for demonstrating the robustness and reproducibility of deep learning models. As presented in Table 2**,** the proposed hybrid framework was evaluated using five different random seeds. The small standard deviations indicate that the reported performance is not strongly dependent on a particular random initialization, supporting the consistency of the proposed framework.

Table 2. Statistical Validation of the Proposed Hybrid Framework Across Multiple Random Seeds

Seed	Accuracy (%)	EER (%)
40	98.82	1.18
41	98.89	1.09
42	98.95	1.02
43	98.91	1.06
44	98.87	1.12

The average performance across five independent runs was

- Mean Accuracy = 98.89 ± 0.05%
- Mean EER = 1.09 ± 0.06%

The small standard deviation demonstrates that the proposed framework produces consistent and reliable performance across multiple training runs, indicating good robustness and reproducibility.

4.8 Discussion

The experimental results demonstrate that combining self-supervised Wav2Vec2 embeddings with handcrafted LFCC features improves spoof detection performance compared with using either representation individually. Wav2Vec2 captures contextual characteristics from the raw speech waveform, whereas LFCC features capture spectral characteristics that can reflect artifacts introduced during speech synthesis and voice conversion. The ablation results provide direct evidence of this complementary behaviour. The hybrid model achieves 98.95% accuracy and an EER of 1.02%, compared with 97.80% accuracy and 1.80% EER for Wav2Vec2 and 96.90% accuracy and 2.10% EER for LFCC-CNN. The improvement over both individual branches indicate that the fusion provides additional information that is useful for distinguishing bona fide and spoofed speech.

The classification ability of the proposed framework is further verified by the confusion matrix and ROC analysis. From the confusion matrix we observe that the majority of the bona fide and spoofed utterances are classified correctly, and very few samples are misclassified. Alternatively, the ROC curve provides a way to view the model's discrimination ability across different decision thresholds. Moreover, results from five different random seeds are presented, and they show small variations in accuracy and EER. The mean accuracy of 98.89% and the mean EER of 1.09% demonstrate consistent performance for the performed runs.

The comparison with existing methods also shows that the proposed framework achieves competitive performance on the ASVspoof2019 LA benchmark. However, the evaluation is limited to a single benchmark dataset, and performance under unseen attack types or cross-dataset conditions may differ. Future work can therefore investigate cross-corpus evaluation using datasets such as ASVspoof2021 LA/DF and In-the-Wild, together with lightweight model designs for real-time deployment.

5. CONCLUSIONS

The covert attacks of voice spoofing that are created with the help of hi-tech text-to-speech (TTS) and voice conversion (VC) technologies pose a serious threat to the accuracy of the Automatic Speaker Verification (ASV) system. This paper introduced a hybrid voice anti-spoofing system that combines a self-supervised Wav2Vec2 back-end with manually designed Linear Frequency Cepstral Coefficient (LFCC) features to utilize complementary contextual and spectral data to identify spoofing. The Wav2Vec2 line encodes the

high-level semantic features directly out of raw speech signals, and the LFCC-CNN line encodes fine-grained spectral features of synthetic and converted speech. The merged characteristic representation allows better differentiation between bona fide utterances and spoofed utterances and a lightweight classification structure.

The proposed framework was tested on the ASVspoof2019 Logical Access (LA) dataset in accordance with the conventional evaluation procedure accepted in the field of voice anti-spoofing studies. The results of the experimental work showed that it was very well classified with an overall accuracy of 98.95, a precision of 98.96, a recall of 98.94, an F1-score of 98.95, and a low equal error rate (EER) of 1.02. The ablation experiment also supported the claim that the combination of Wav2Vec2 embeddings and LFCC features was always better than the branches of feature extraction, which proved the efficiency of the suggested feature fusion strategy. The proposed framework is also very robust, as it was statistically validated through multiple random seeds, which resulted in consistent and repeatable performance.

The hybrid architecture proposed provides a feasible approach to enhancing biometric security in the applications of financial authentication, intelligent voice assistants, secure access control, and speaker verification systems. The framework can combine contextual speech representations with hand-crafted acoustic features to increase resistance to more recent voice spoofing attacks and still be computationally efficient.

The proposed model was able to perform competitively on the ASVspoof2019 LA benchmark but was only tested on one dataset. Future research will revolve around cross-corpus validation with more challenging data like ASVspoof2021 LA/DF and In-the-Wild, exploration of more advanced attention-based feature fusion methods, optimization to deploy in resource-constrained devices, and greater resistance to emerging deepfake speech generation.

REFERENCES

[1]. Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," Speech Communication, vol. 66, pp. 130–153, 2015, doi: 10.1016/j.specom.2014.10.005.

[2]. M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," Computer Speech & Language, vol. 45, pp. 516–535, 2017, doi: 10.1016/j.csl.2017.01.001.

[3]. Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilçi, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, and H. Delgado, "ASVspoof: The Automatic Speaker Verification Spoofing and Countermeasures Challenge," IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 4, pp. 588–604, Jun. 2017, doi: 10.1109/JSTSP.2017.2671435.

[4]. X. Wang, H. Delgado, H. Tak, J.-w. Jung, H.-j. Shim, M. Todisco, I. Kukanov, X. Liu, M. Sahidullah, T. H. Kinnunen, N. Evans, K. A. Lee, and J. Yamagishi, "ASVspoof 5: Crowdsourced Speech Data, Deepfakes, and Adversarial Attacks at Scale," in Proc. Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024), 2024, pp. 1–8, doi: 10.21437/ASVspoof.2024-1.

[5]. Y. Eom, Y. Lee, J. S. Um, and H. R. Kim, "Anti-Spoofing Using Transfer Learning with Variational Information Bottleneck," in Proc. Interspeech, 2022, pp. 3568–3572, doi: 10.21437/Interspeech.2022-10200.

[6]. H. Tak, J.-W. Jung, J. Patino, M. Todisco, and N. Evans, "Graph Attention Networks for Anti-Spoofing," in Proc. Interspeech, 2021, pp. 2356–2360, doi: 10.21437/Interspeech.2021-993.

[7]. B. Huang, S. Cui, J. Huang, and X. Kang, "Discriminative Frequency Information Learning for End-to-End Speech Anti-Spoofing," IEEE Signal Processing Letters, vol. 30, pp. 185–189, 2023, doi: 10.1109/LSP.2023.3251895.

[8]. Y. El Kheir, A. Das, E. E. Erdogan, F. Ritter-Guttierez, T. Polzehl, and S. Möller, "Two Views, One Truth: Spectral and Self-Supervised Features Fusion for Robust Speech Deepfake Detection," in Proc. 2025 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2025, pp. 1–5, doi: 10.1109/WASPAA66052.2025.11230938.

[9]. A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, 2020, pp. 12449–12460, doi: 10.48550/arXiv.2006.11477.

[10]. M. J. Alam, P. Kenny, G. Bhattacharya, and T. Stafylakis, "Development of CRIM system for the automatic speaker verification spoofing and countermeasures challenge 2015," in Proc. Interspeech, 2015, pp. 2072–2076, doi: 10.21437/Interspeech.2015-469.

[11]. X. Wang et al., "ASVspoof 2019: A Large-Scale Public Database of Synthesized, Converted and Replayed Speech," Computer Speech & Language, vol. 64, Art. no. 101114, 2020, doi: 10.1016/j.csl.2020.101114.

[12]. J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, and H. Delgado, "ASVspoof 2021: Accelerating Progress in Spoofed and Deepfake Speech Detection," in Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge (ASVspoof 2021), 2021, pp. 47–54, doi: 10.21437/ASVSPOOF.2021-8.

[13]. X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, and K. A. Lee, "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, doi: 10.1109/TASLP.2023.3285283.

[14]. M. Neelima and I. S. Prabha, "Hybrid feature optimization for voice spoof detection using CNN-LSTM," Traitement du Signal, vol. 41, no. 2, pp. 717–727, Apr. 2024, doi: 10.18280/ts.410214.

[15]. J. Zhou, H. Tao, D. J. Jawawi, D. Wang, E. Ibeke, and C. Biamba, "Voice Spoofing Countermeasure for Voice Replay Attacks Using Deep Learning," Research Square preprint, Jul. 2022, doi: 10.1186/s13677-022-00306-5.

[16]. T. Arif, A. Javed, M. Alhameed, F. Jeribi, and A. Tahir, "Voice Spoofing Countermeasure for Logical Access Attacks Detection," IEEE Access, vol. 9, pp. 162857–162868, 2021, doi: 10.1109/ACCESS.2021.3133134.

[17]. K. Shahzad, S. Farhan, Yasin-Ul-Haq, R. Sana, and S. Pathan, "Enhancing Voice Spoofing Detection: A Hybrid Approach with VGGish-LSTM Model for Improved Security in Automatic Speaker Verification Systems," IEEE Access, vol. 13, pp. 40682–40702, 2025, doi: 10.1109/ACCESS.2025.3544639.

[18]. Y. Ren, H. Peng, L. Li, and Y. Yang, "Lightweight Voice Spoofing Detection Using Improved One-Class Learning and Knowledge Distillation," IEEE Transactions on Multimedia, vol. 26, pp. 4360–4374, 2024, doi: 10.1109/TMM.2023.3321505.

[19]. Z. M. Almutairi and H. Elgibreen, "Detecting Fake Audio of Arabic Speakers Using Self-Supervised Deep Learning," IEEE Access, vol. 11, pp. 72134–72147, 2023, doi: 10.1109/ACCESS.2023.3286864.

[20]. Y. Lee, N. Kim, J. Jeong, and I.-Y. Kwak, "Experimental Case Study of Self-Supervised Learning for Voice Spoofing Detection," IEEE Access, vol. 11, pp. 24216–24226, 2023, doi: 10.1109/ACCESS.2023.3254880.

[21]. J. M. Martín-Doñas and A. Álvarez, "The Vicomtech Audio Deepfake Detection System Based on Wav2Vec2 for the 2022 ADD Challenge," arXiv preprint arXiv:2203.01573, 2022, doi: 10.48550/arXiv.2203.01573.

[22]. Y. Xiao and R. K. Das, "XLSR-Mamba: A Dual-Column Bidirectional State Space Model for Spoofing Attack Detection," IEEE Signal Processing Letters, vol. 32, pp. 1276–1280, 2025, doi: 10.1109/LSP.2025.3547861.

[23]. F. Javanmardi, S. R. Kadiri, and P. Alku, "Exploring the Impact of Fine-Tuning the Wav2vec2 Model in Database-Independent Detection of Dysarthric Speech," IEEE Journal of Biomedical and Health Informatics, vol. 28, no. 8, pp. 4951–4962, Aug. 2024, doi: 10.1109/JBHI.2024.3392829.

[24]. C. Yi, J. Wang, N. Cheng, S. Zhou, and B. Xu, "Applying Wav2Vec2.0 to Speech Recognition in Various Low-Resource Languages," arXiv preprint arXiv:2012.12121, Dec. 2020, doi: 10.48550/arXiv.2012.12121.

[25]. I.-Y. Kwak, S. Kwag, J. Lee, Y. Jeon, J. Hwang, H.-J. Choi, J.-H. Yang, S.-Y. Han, J. H. Huh, C.-H. Lee, and J. W. Yoon, "Voice Spoofing Detection Through Residual Network, Max Feature Map, and Depthwise Separable Convolution," IEEE Access, vol. 11, pp. 49140–49152, 2023, doi: 10.1109/ACCESS.2023.3275790.

[26]. B. Ustubioglu, G. Tahaoglu, A. Ustubioglu, G. Ulutas, I. Amerini, and M. Kilic, "Multi Pattern Features-Based Spoofing Detection Mechanism Using One Class Learning," IEEE Access, vol. 12, pp. 117523–117540, 2024, doi: 10.1109/ACCESS.2024.3447572.

[27]. Y. Zhang, Z. Li, J. Lu, W. Wang, and P. Zhang, "Synthetic Speech Detection Based on the Temporal Consistency of Speaker Features," IEEE Signal Processing Letters, vol. 31, pp. 944–948, 2024, doi: 10.1109/LSP.2024.3381890.

[28]. L. Zhang, X. Wang, E. Cooper, N. Evans, and J. Yamagishi, "The PartialSpoof Database and Countermeasures for the Detection of Short Fake Speech Segments Embedded in an Utterance," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 813–825, 2023, doi: 10.1109/TASLP.2022.3233236.

[29]. Z. Wang and J. H. L. Hansen, "Toward Improving Synthetic Audio Spoofing Detection Robustness via Meta-Learning and Disentangled Training with Adversarial Examples," IEEE Access, vol. 12, pp. 99894–99911, 2024, doi: 10.1109/ACCESS.2024.3421281.

[30]. O. A. Shaaban, R. Yildirim, and A. A. Alguttar, "Audio Deepfake Approaches," IEEE Access, vol. 11, pp. 132652–132682, 2023, doi: 10.1109/ACCESS.2023.3333866.

[31]. J. Lu, Y. Zhang, Z. Li, Z. Shang, W. Wang, and P. Zhang, "Leveraging Distance Information for Generalized Spoofing Speech Detection," Computer Speech & Language, vol. 94, Art. no. 101804, 2025, doi: 10.1016/j.csl.2025.101804.

[32]. T. Kinnunen, H. Delgado, N. Evans, K. A. Lee, V. Vestman, A. Nautsch, M. Todisco, X. Wang, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, "Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification: Fundamentals," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2195–2210, 2020, doi: 10.1109/TASLP.2020.3009494.

[33]. Z. Akhtar, T. L. Pendyala, and V. S. Athmakuri, "Video and Audio Deepfake Datasets and Open Issues in Deepfake Technology: Being Ahead of the Curve," Forensic Sciences, vol. 4, no. 3, pp. 289–377, 2024, doi: 10.3390/forensicsci4030021.

[34]. M. Sharafudeen et al., "A Blended Framework for Audio Spoof Detection with Sequential Models and Bags of Auditory Bites," Scientific Reports, vol. 14, Art. no. 20192, 2024, doi: 10.1038/s41598-024-71026-w.