Robust Deepfake Audio Detection Using Multi-Branch Feature Fusion and Adversarial Hardening

Simrithaa N V[1], Royo J Varghese[2], Mithun Karthik[3], Shri Kirthikha Gunasekaran[4], Surya R[5]
[1,2,3,4,5]*Students, Bachelor of Engineering Computer Science, PSG College of Technology*

***Abstract*—Rapid progress in neural text-to-speech and voice conversion has made synthetic speech almost indistinguishable from genuine human speech, creating serious risks for speaker verification, digital forensics, and financial security. Although recent spectro-temporal and convolutional neural network (CNN) detectors achieve high accuracy on benchmark datasets, they remain vulnerable to adversarial perturbations and generalize poorly to real-world audio. This paper presents a robust deepfake speech detector that combines three complementary branches interpretable classical features, an EfficientNet-based CNN operating on log-Mel spectrograms, and self-supervised learning (SSL) embeddings fused at the score level. The CNN branch is hardened using Fast Gradient Sign Method (FGSM) adversarial training, and domain-adaptive augmentation is applied during training to improve robustness under codec, channel, and noise variation. The system is trained and evaluated on the ASVspoof 2021 Logical Access dataset for both binary detection and multi-class spoof attribution. The fused model achieves 99.30% binary accuracy and an Equal Error Rate (EER) of 0.97%, together with 94.20% multi-class attribution accuracy, outperforming the individual branches and the spectro-temporal baseline while remaining stable under adversarial attack.**

***Index Terms*—Adversarial training, deepfake audio detection, feature fusion, self-supervised learning, spoofing countermeasures.**

I. INTRODUCTION

The current pace of progress in artificial intelligence and deep learning has driven remarkable improvements in speech-synthesis technologies, including neural text-to-speech (TTS) systems and voice conversion (VC) methods. These systems now produce audio that is almost indistinguishable from a real human voice. While such capabilities enable positive applications such as virtual assistants and accessibility tools, they also raise serious security concerns. The most pressing of these is the use of deepfake audio synthetic speech generated to impersonate a target speaker.

Deepfake audio poses a significant threat to speaker verification systems, digital forensics, financial security, and the detection of misinformation. Malicious actors can exploit these technologies to bypass authentication systems, commit fraud, or spread false information. Consequently, developing effective deepfake audio detection systems has become essential.

Recent detection methods rely on classical signal-processing techniques and deep learning models. In particular, spectro-temporal frameworks combined with convolutional neural network (CNN) encodings have achieved strong results. However, despite high accuracy on benchmark datasets such as the ASVspoof 2021 Logical Access (LA) dataset, most existing systems struggle under realistic, unconstrained conditions.

Two problems are especially important. First, deep learning models are susceptible to adversarial attacks, in which carefully crafted perturbations mislead the model without any audible degradation. Second, variations in codec, transmission channel, and environmental noise can severely degrade detection performance. These issues highlight the need for detectors that are not only accurate but also resistant to adversarial manipulation and robust to domain shift.

To address these challenges, this paper presents a robust deepfake speech detector that combines multi-branch feature fusion with adversarial training. The system integrates classical audio features, a CNN-based spectrogram representation, and self-supervised learning (SSL) embeddings to capture complementary information from the speech signal. The Fast Gradient Sign Method (FGSM) is used for adversarial

hardening to improve resistance to perturbation-based attacks. The proposed methodology aims to improve detection accuracy, robustness, and generalization simultaneously.

1.1 Problem Statement

The growing sophistication of deepfake audio produced by neural TTS and VC tools presents a serious threat to current audio-authentication mechanisms. Although existing detection models achieve high accuracy on benchmark datasets such as the ASVspoof 2019 LA and ASVspoof 2021 LA datasets, they tend to fail when applied to real-world scenarios. These models are particularly vulnerable to adversarial attacks, in which even small perturbations can confuse the model without degrading perceptual quality. Moreover, their reliance on fine spectral characteristics limits their ability to extract deeper semantic information, so more robust and generalized detection methods are required.

1.2 Current Challenges

- Detectors can be fooled by small adversarial perturbations that are imperceptible to humans, leaving systems exposed to real-world attacks.
- Models that are highly accurate on benchmark datasets such as ASVspoof 2019 LA often do not sustain that accuracy on real-world audio variations.
- Many current methods rely primarily on surface-level spectral features and cannot capture deeper semantic and contextual meaning in speech.
- Poor generalization across conditions -background noise, recording devices, and compression effects - reduces reliability in real-life settings.
- Heavy reliance on CNN-based architectures narrows the variety of learned features and can miss complementary representations.
- The absence of strong training techniques such as adversarial training makes models less resistant to unseen attacks and weakens overall system security.

1.3 Proposed Solution

The proposed solution is a robust multi-branch deepfake speech detector that exploits complementary feature representations to improve reliability and generalization. The system processes the input audio through three parallel branches: a classical feature branch, which extracts interpretable spectro-temporal features that provide a stable basis for detecting spoofing artifacts; a CNN-based spectrogram branch, built on EfficientNet, which detects fine-grained artifacts introduced by synthesis methods; and a self-supervised learning (SSL) branch, which models high-level semantic and contextual information. The CNN branch is explicitly trained to identify discriminative patterns in log-Mel spectrograms, while the SSL embeddings improve extrapolation to unseen spoofing methods. Each branch produces an independent prediction score, and these are combined through score-level fusion to reach the final decision. This heterogeneous design ensures that the strengths of one branch compensate for the weaknesses of others, balancing interpretability, robustness, and detection performance across diverse conditions.

During training, the framework applies adversarial training to increase robustness. Gradient-based adversarial perturbations generated with the Fast Gradient Sign Method (FGSM) are used to selectively harden the CNN branch. By combining adversarial hardening and higher-order feature fusion in a single pipeline, the proposed system seeks to improve reliability and effectiveness in both controlled and real-world settings.

1.4 Scope

This work develops a robust deepfake audio detector based on multi-branch feature fusion and adversarial training. It uses the benchmark ASVspoof 2021 LA dataset for training and evaluation. The focus is on improving detection accuracy, resilience to adversarial examples, and the ability to generalize across diverse audio settings, while retaining the computational efficiency and interpretability needed for practical deployment in real-world security systems.

II. LITERATURE SURVEY

This section reviews prior work on deepfake audio detection, CNN-based models, and adversarial robustness. It summarizes existing approaches and their limitations - such as vulnerability to attacks and limited generalization and highlights the need for more robust detection systems.

2.1 Review of Related Work

Can and Soyhan (2025) propose a fusion-based method that improves deepfake speech detection by combining spectro-temporal features with CNN embeddings, and introduce the concept of spoof-system attribution to identify attack sources. Their model performs well on benchmark datasets by using both interpretable and learned features. However, it does not assess adversarial robustness and is not grounded in semantic or contextual evaluation of speech.

Rabhi et al. (2024) present a detailed study of adversarial attacks on audio deepfake detection systems, discussing a range of attack strategies and possible defenses. The work highlights the fragility of current detection models under adversarial settings. The proposed defenses, however, are computationally expensive, do not scale well, and do not provide a single robust detection architecture.

Kwon and Nam (2024) propose a classification-score-based approach for detecting adversarial audio in speech-recognition systems that operates without modifying the underlying automatic speech recognition (ASR) model. The method is lightweight and model-agnostic, making it easy to integrate. It is, however, primarily concerned with securing the ASR system and does not address deepfake speech detection or spoofing attacks such as TTS and VC.

Asuail et al. (2025) propose a hybrid CNN–LSTM model that combines Mel-frequency cepstral coefficients (MFCC) and spectrogram features to improve audio deepfake detection over CNN-only models. The hybrid architecture yields a richer feature representation, but it provides no adversarial robustness, is tested on relatively small datasets, and lacks interpretability mechanisms.

Hashmi et al. (2025) present a human-cognition-inspired ensemble of audio and visual transformers for detecting video deepfakes, using cross-modal consistency between audio and visual cues. The model is designed for combined audio–visual input rather than standalone audio, and its high computational cost makes real-time deepfake audio detection impractical.

Kaur et al. (2024) introduce a fusion model that integrates several spectral features to improve audio deepfake detectability. Their CNN-based ensemble achieves higher accuracy than single-feature methods but does not include an adversarial-robustness evaluation, is run on small datasets, and does not address generalization.

Bohara and Bairwa (2025) detect deepfake audio by comparing models trained on spectrogram-based features, confirming that time–frequency representations are effective for distinguishing synthetic from authentic audio. The authors caution that classical machine-learning methods generalize poorly, have not been tested under adversarial hardening or anti-spoofing conditions, and have received limited attention with respect to model-development strategy.

Wang et al. (2025) develop a multimodal (audio–visual) model for deepfake detection that exploits complementary modalities to improve robustness. The approach does not adequately address audio-only spoofing, is computationally complex, and has practical limitations.

Wani and Amerini (2025) present a transformer-based dynamic knowledge-condensation model with an audio selective transformer that emphasizes the most relevant audio features for efficient deepfake detection. The model performs well through selective feature filtering, but its limitations include sensitivity to dataset complexity, a lack of detailed adversarial evaluation, and low cross-dataset generalization.

2.2 Research Gap

- Limited resistance to adversarial attacks in current deepfake audio detection systems.
- Minimal use of self-supervised learning (SSL) to capture semantic and contextual features of speech.
- Weak generalization to other datasets and to real-world audio scenarios.
- Little integration across different feature types (spectro-temporal, CNN, and semantic features).
- High computational complexity of advanced transformer and multimodal models.
- A lack of balanced approaches that provide both interpretability and high detection performance.

2.3 Inference

The literature shows that substantial progress has been made in deepfake audio detection through spectro-temporal features, CNN-based architectures, and hybrid designs. Although these techniques achieve high accuracy on test data, they are often not robust in

practice. A key issue is their susceptibility to adversarial attacks, which can mislead models with only small perturbations. In addition, most techniques do not capture deeper semantic characteristics and do not generalize well across settings. An improved, more robust, and adaptive framework that addresses these limitations is therefore desirable.

III. METHODOLOGY

The system is designed as a robust, multi-branch deepfake speech detection platform that addresses real-world problems such as adversarial attacks through a structured pipeline. The role of each component feature extraction, model training, and decision fusion is clearly defined. The system is organized as autonomous modules that communicate in well-defined stages, making it scalable, extensible, and easy to test. By combining classical signal-processing features, deep-learning spectrogram analysis, and self-supervised speech embeddings, the system captures both low-level and high-level properties of the audio signal, aiming to be accurate, interpretable, and efficient for practical use.

3.1 Overall Architecture

The proposed system uses a multi-branch structure to improve the accuracy, robustness, and generalization of deepfake audio detection. Rather than relying on a single model, the input audio is processed by three parallel branches, each of which captures different attributes of the signal. As shown in Figure 1, the pipeline begins with input audio samples drawn from the benchmark dataset. These signals are first pre-processed to ensure consistency and quality, and the audio is then routed in parallel into three feature-extraction branches. The first branch derives classical, handcrafted features such as MFCC, spectral, and energy-based descriptors. The second branch converts audio into Mel spectrograms and processes them with a convolutional neural network, and is further hardened using adversarial training. The third branch employs self-supervised learning models to derive deep semantic embeddings from raw audio. Each branch produces an independent prediction score, and the scores are fused at the score level via weighted averaging; the classification module then labels the input audio as either genuine (bonafide) or spoofed.

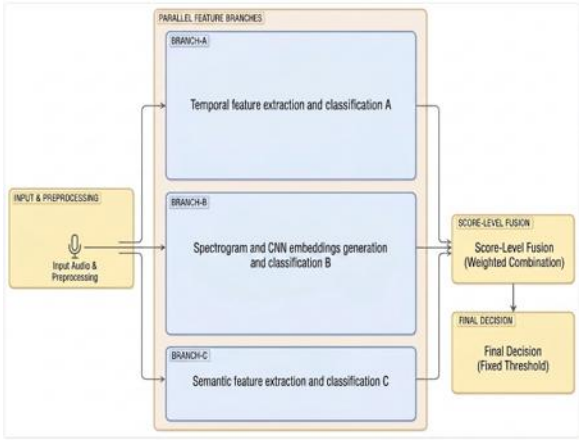

Figure 1. Overall architecture of the proposed multi-branch detector.

3.2 Dataset

The system is developed and evaluated on the benchmark ASVspoof 2021 LA dataset, which is widely used in deepfake audio detection research. This dataset contains real (bonafide) and spoofed speech samples produced with a range of TTS and VC methods. It follows the default ASVspoof2021-LA protocol, with realistic codec and channel distortions and thirteen distinct spoofing systems (A07 through A19). These systems span a variety of synthesis approaches, including neural TTS models (for example, Tacotron-based systems with WaveNet/WaveRNN vocoders), traditional vocoder-based methods (for example, WORLD and Griffin-Lim), hybrid TTS-to-VC pipelines, and voice-conversion methods based on autoencoders and embedding-based mappings. The spoofing systems are summarized in Table 1.

Including many spoofing systems introduces high variability that simulates real-world conditions, in which several synthesis and conversion methods may be encountered. This range enables both training and testing across many potential attack types, improving robustness when the model encounters unfamiliar or difficult audio in practice.

Table 1. Spoof Types of the ASV spoof 2021 LA Dataset

System	Category	Acoustic / Conversion Core	Vocoder / Synthesis	Definition
A07	TTS	LSTM acoustic model (WORLD vocoder front end)	WORLD + WaveCycleGAN2 post-filter	WORLD output refined by neural post-filter for higher naturalness
A08	TTS	Neural source-filter TTS	Neural source-filter (no classic vocoder)	Fast, high-quality parametric generation
A09	TTS	SPSS with separate LSTMs (duration and acoustics)	Vocaine	Lightweight, real-time oriented method
A10	TTS	Tacotron 2 (seq2seq acoustic front end)	WaveRNN	Strong naturalness and speaker similarity
A11	TTS	Tacotron 2 front end	Griffin-Lim	Faster reconstruction; trades some quality versus neural vocoders
A12	TTS	SPSS-conditioned prosodic features	WaveNet (autoregressive)	Prosody-conditioned autoregressive waveform generation
A13	Hybrid (TTS to VC)	Moment-matching conversion	Differential spectral filtering at the waveform level	Converts TTS output toward target speaker
A14	Hybrid (TTS to VC)	ASR bottleneck features feeding an LSTM converter	STRAIGHT	Bottleneck-driven conversion; STRAIGHT resynthesis
A15	Hybrid (TTS to VC)	Similar to A14 conversion	WaveNet	Hybrid conversion with neural vocoder rendering
A16	TTS (concatenative)	Unit-selection stitching of recorded units	Concatenative synthesis	Classic baseline / reference for traditional TTS
A17	VC	Variational autoencoder (feature-space conversion, e.g., MCCs)	Waveform filtering / MLSA	VAE-based spectral conversion from bonafide recordings
A18	VC	i-vector / PLDA embedding prediction; MFCC conversion	Dedicated vocoder	Embedding-space mapping to target speaker
A19	VC	Learned spectral transfer function	Spectral transfer synthesis	Applies target spectral characteristics to source

3.3 Data Preparation and Augmentation

The metadata is first read and filtered to retain the bonafide and spoofed classes, and each audio clip (stored in .flac format) is linked to its metadata through a unique trial identifier. The dataset is then split into training, validation, and evaluation sets. To enrich learning, data augmentation additive noise, reverberation, and channel simulation is applied to emulate real-world conditions and improve model robustness.

3.4 Audio Preprocessing

The preprocessing stage normalizes the input audio. All samples are resampled to a fixed rate of 16 kHz. This standardizes the input across samples and improve the quality of the subsequent feature-extraction stages.

3.5 Classical Feature Branch

This branch extracts handcrafted features from the audio signal, including MFCC, delta MFCC, chroma, spectral contrast, tonnetz. These are assembled into a fixed 51-dimensional feature vector for each sample, which is then used to train a machine-learning classifier such as XGBoost for binary classification. The approach is simple and efficient and provides a strong baseline for deepfake detection based on low-level statistical features.

3.6 CNN Spectrogram Branch

The CNN branch operates on a time–frequency representation of the audio. Each signal is converted into a log-Mel spectrogram, stored as a numerical array, and normalized to ensure balanced input. The spectrogram is passed through a convolutional neural network with several convolutional layers followed by activation and pooling layers, and fully connected layers produce the probability of the audio being real or spoofed. Because it learns spatial patterns in the spectrogram, this branch is particularly effective at detecting artifacts present in synthetic audio. The branch is further strengthened with adversarial training, as detailed in Section 3.7.

3.7 Adversarial Training of the CNN Branch

Among the three branches, the CNN spectrogram branch is the only component that is differentiable from input to output, and it is therefore the branch that is explicitly hardened against adversarial manipulation. Because the log-Mel spectrogram is fed directly into a convolutional network whose fully connected layers emit the spoof probability, gradients can be propagated all the way back to the spectrogram itself. This property makes the branch an ideal target for gradient-based adversarial training: the very mechanism an attacker would exploit to fool the model is instead used, during training, to make the model resistant to it.

The perturbation is generated using the Fast Gradient Sign Method (FGSM). Given a clean spectrogram x with true label y and network parameters θ, an adversarial spectrogram is obtained in a single step as

$$x_{\text{adv}} = x + \varepsilon \cdot \text{sign}\big(\nabla_x L(\theta, x, y)\big) \quad (1)$$

where L is the classification loss and ε controls the magnitude of the perturbation. The defining characteristic of this perturbation is that it is not random noise. The sign of the input gradient identifies, for every time–frequency bin, the direction in which a small change most increases the loss, so the added signal is deliberately aligned with the model's weakest direction. Keeping ε small ensures the perturbation stays imperceptible while remaining maximally disruptive to the classifier. An iterative extension such as Projected Gradient Descent (PGD) applies this step repeatedly for a stronger attack; the single-step FGSM formulation is adopted here for its low computational cost, which allows adversarial examples to be generated on the fly during every training iteration.

Adversarial training then augments each mini-batch with these crafted examples. For every batch, the model first performs a forward pass on the clean spectrogram and computes the clean loss. The gradient of this loss with respect to the input is used to construct the adversarial spectrogram according to Equation (1), and a second forward pass is performed on the perturbed input to obtain the adversarial loss. The two losses are combined into a single objective,

$$L_{\text{total}} = L_{\text{clean}} + 0.5 \cdot L_{\text{adv}} \quad (2)$$

and the network is updated with a single backward pass on this combined loss. Training on clean and adversarial samples simultaneously forces the network to learn features that stay discriminative even when the input has been deliberately shifted toward misclassification. The weight of 0.5 on the adversarial term balances the two objectives, so the model gains robustness to perturbed inputs without sacrificing accuracy on clean speech. The full training loop is summarized in Figure 2.

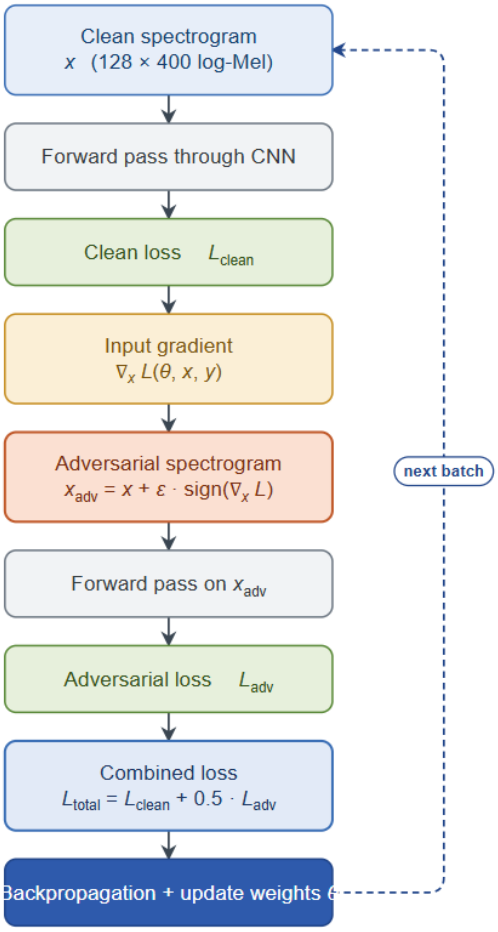

Figure 2. FGSM adversarial-training loop for the CNN branch.

An important design choice is that the perturbation is applied at the spectrogram level rather than to the raw waveform or to the extracted feature vectors. In principle the attack could be carried out on the raw audio, since gradients do flow back through the spectrogram transformation to the waveform; in practice this path is far less stable, because small changes to the waveform translate into unpredictable changes in the time–frequency representation, making the resulting perturbations difficult to control and to interpret. Operating directly on the spectrogram avoids this instability: it is a structured, fixed time–frequency

representation that matches the network input exactly and yields well-behaved gradients. The compressed feature vectors used by the classical and SSL branches, by contrast, cannot be attacked in this manner at all. Those features are consumed by tree-based classifiers such as XGBoost, which are non-differentiable, so no gradient path of the form $\partial L/\partial(\text{feature})$ exists. Furthermore, an arbitrary edit to an element of a compressed feature vector has no correspondence to any realizable change in the underlying audio, rendering such a perturbation mathematically possible but physically meaningless. For these reasons, adversarial hardening is applied selectively to the CNN branch, where it is both technically well-defined and physically interpretable.

To support efficient batched adversarial training, all spectrograms are standardized to a fixed shape of 128 × 400 (128 Mel bands × 400 time frames) prior to training, with longer clips truncated and shorter clips padded. This uniform input size allows clean and adversarial batches to be processed together and keeps the on-the-fly FGSM step computationally light. As reported in Section IV, this procedure yields a CNN branch that retains high accuracy on clean audio while remaining stable under FGSM attack, in contrast to the plain CNN, whose accuracy collapses under the same perturbation.

3.8 SSL Embedding Branch

The SSL branch uses pretrained self-supervised models such as wav2vec 2.0 to extract deep features directly from the raw audio waveform. The model receives raw audio and produces a high-dimensional representation that captures semantic and contextual information. These embeddings are passed to a classifier like XGBoost to label the sample as authentic or spoofed. By capturing high-level speech properties, this branch improves the generalization capability of the overall system.

3.9 Score-Level Fusion

The outputs of the classical, CNN, and SSL branches are combined at the score level. Each branch produces a prediction score corresponding to the probability that the audio is spoofed, and these scores are merged using a weighted average. The weight assigned to each branch is determined on validation data to obtain the best performance. This strategy allows the system to benefit from the complementary strengths of the different feature representations, yielding more accurate results.

3.10 Classification and Decision

The final decision is made from the fused score using a threshold-based rule. If the fused score exceeds the threshold, the input is classified as spoofed; otherwise, it is classified as genuine. The optimal threshold is selected on the validation set to balance false-positive and false-negative rates, supporting reliable and robust predictions.

3.11 System Workflow

The end-to-end workflow takes raw audio as input and outputs a prediction. It begins with raw audio in .flac format, which is pre-processed through resampling, normalization, and noise reduction so that all branches receive standardized input. The audio is then sent through the three branches in parallel: classical features are extracted and classified with XGBoost; the log-Mel spectrogram is fed into the adversarially trained CNN; and the raw waveform is processed by an SSL model such as wav2vec 2.0 to produce deep embeddings that are then classified. The three scores are fused and thresholder to produce the final Bonafide /spoof decision, as summarized in Figure 3.

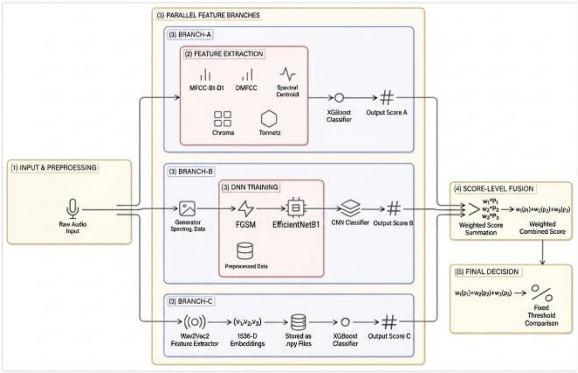

Figure 3. End-to-end system workflow.

3.12 Training and Evaluation Setup

The models are trained on the training set and assessed on the development set. Optimization algorithms such as Adam are used for stable convergence. System performance is measured using metrics including accuracy and F1-score, and robustness is additionally analyzed on adversarial samples. This systematic training and evaluation procedure ensures effective operation under a range of conditions.

IV. RESULTS

This section presents the experimental findings for the proposed deepfake audio detection model. A stage-by-stage evaluation methodology is followed to ensure a fair comparison. All models are evaluated on the ASVspoof 2021 LA dataset using identical data-preparation and feature-extraction procedures. The assessment consists of two stages: binary classification (bonafide versus spoof) and multi-class spoof attribution (A07–A19).

4.1 Evaluation Protocol

All models use the same preprocessing steps resampling, normalization, and feature extraction and the evaluation is divided into two tasks. Stage 1 performs binary classification to separate bonafide and spoofed audio, and Stage 2 performs multi-class classification to distinguish among spoofing attack types. This design supports an unbiased and fair comparison across all models.

4.2 Models Under Comparison

Both baseline and proposed models are presented within a single pipeline. The baseline models are a temporal model using a handcrafted 51-dimensional feature set with XGBoost, and a CNN model operating on spectrograms. The proposed models extend this baseline with an SSL model based on wav2vec embeddings, an adversarially trained CNN for robustness, and a final fusion model that aggregates all branches. This progression demonstrates the contribution of each component.

4.3 Binary Classification Performance

In the first stage, the models are evaluated for binary classification as onafide (real) or spoofed, using accuracy, per-class F1-scores, and confusion matrices. The SSL model shows strong semantic representation, achieving 98.51% binary accuracy, and the adversarially trained CNN reaches 98.67% with good generalization. The proposed fusion model attains the highest performance, with 99.30% accuracy and an F1-score of 0.9961 for spoof detection. The temporal model achieves high accuracy on spoofed speech but produces many false negatives on onafide speech, indicating poor generalization of handcrafted features.

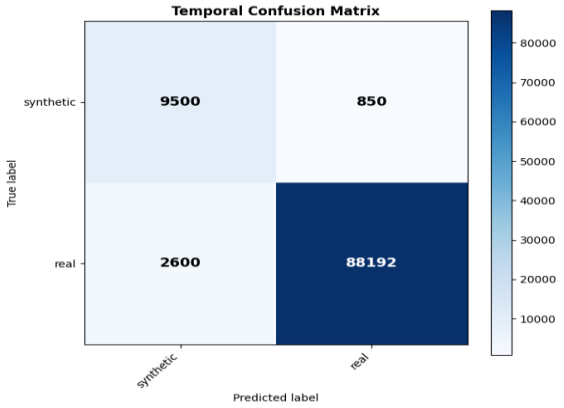

Figure 4. Confusion matrix temporal model.

As shown in Figure 4, the temporal model detects spoofed samples accurately but has a relatively high false-negative rate for bonafide speech, reflecting the limited generalization of handcrafted features.

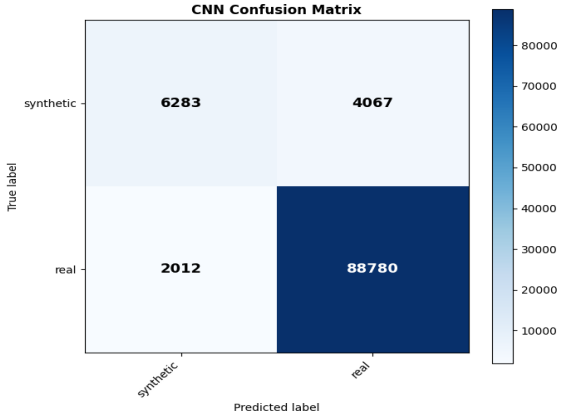

Figure 5. Confusion matrix CNN model.

As shown in Figure 5, the CNN model recognizes bonafide data well but makes comparatively more spoof-recognition errors, revealing sensitivity to spectral variation.

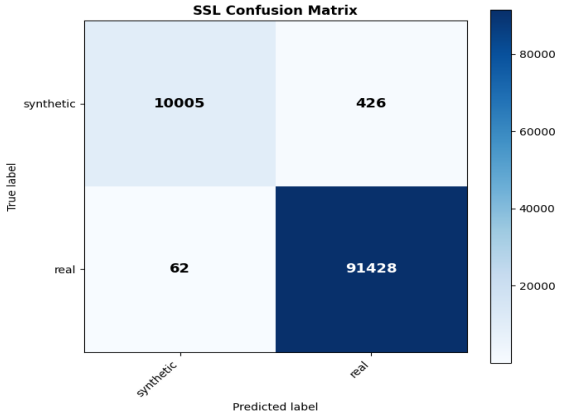

Figure 6. Confusion matrix SSL model.

In contrast, the SSL model in Figure 6 shows clear improvements in detecting both spoofed and bonafide data, with lower error rates driven by stronger semantic generalization.

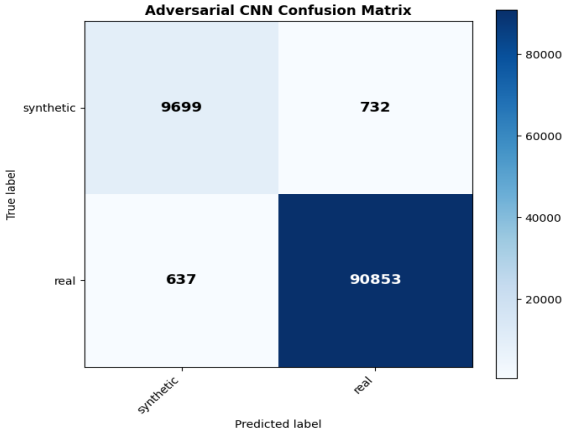

Figure 7. Confusion matrix adversarially trained CNN.

As shown in Figure 7, adversarial training reduces misclassification under perturbation while preserving accurate classification of clean samples.

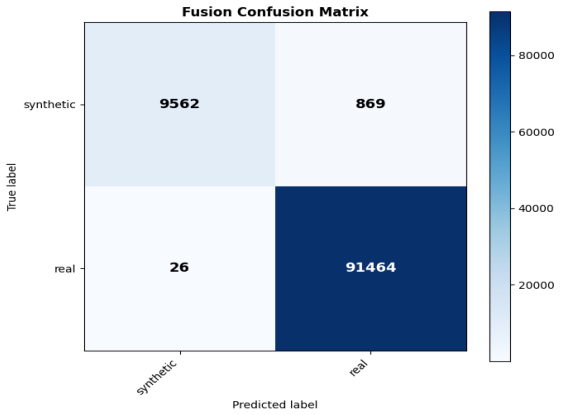

Figure 8. Confusion matrix fusion model.

The fusion model in Figure 8 performs best overall, reducing both false positives and false negatives by combining the three branches.

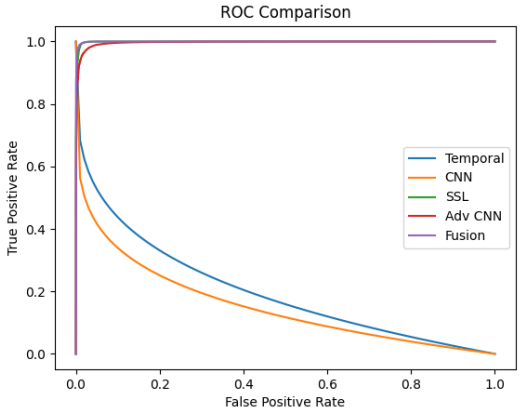

Figure 9. ROC-curve comparison across models.

As shown by the ROC curves in Figure 9, the fusion model maintains a high true-positive rate at any given false-positive rate, while the SSL and adversarial CNN models also show strong discriminative power and the temporal and plain CNN models perform less well. A complete summary of binary and multi-class metrics is given in Table 3.

4.4 Multi-Class Spoof Attribution

In the second stage, the models are evaluated on spoofed samples to determine the type of spoofing attack, using multi-class accuracy together with macro- and weighted-F1 scores. Again the fusion model performs best, with a multi-class accuracy of 94.20% and a macro-F1 of 0.9445. Macro-F1 averages precision and recall equally across all spoof categories, while weighted-F1 accounts for class imbalance. The corresponding spoof-attribution confusion matrix is shown in Figure 10.

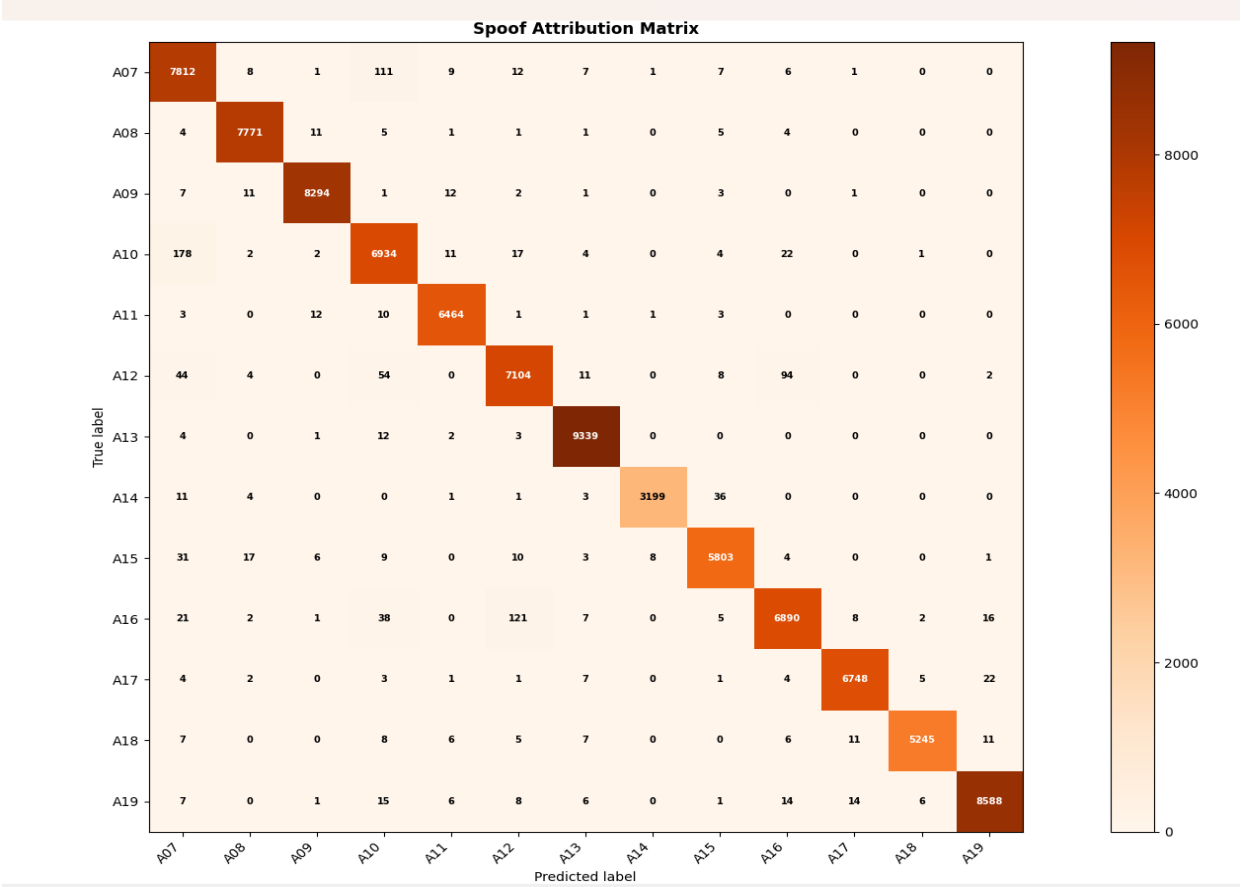

Figure 10. Spoof-attribution confusion matrix (A07–A19).

The attribution matrix in Figure 10 exhibits a strong diagonal, reflecting good classification of most spoofing attacks (A07–A19); some confusion occurs between acoustically similar attacks, but performance is otherwise stable.

4.5 Comparative Performance Analysis

An aggregate comparison of all models reveals a clear improvement in performance as branches are added. The temporal model is a lightweight baseline but lacks representational power. The CNN model improves performance by exploiting spectral information, though it remains vulnerable to adversarial attacks. The SSL model further boosts generalization using high-level features. These results confirm the distinct contribution of each component, as summarized in Figure 11.

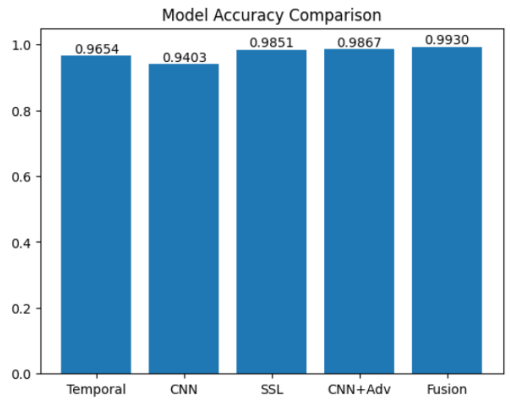

Figure 11. Model-accuracy comparison.

4.6 Adversarial Robustness Analysis

To test robustness, adversarial perturbations are generated with the Fast Gradient Sign Method (FGSM) and added to the spectrograms. The plain CNN model suffers a dramatic drop in accuracy under attack, highlighting its vulnerability, whereas the adversarially trained CNN maintains stable

predictions with only a small performance drop. The fusion model further improves robustness by leveraging predictions from multiple branches to reduce the effect of the attack.

4.7 Equal Error Rate (EER) Analysis

The Equal Error Rate (EER) the operating point at which the false-acceptance and false-rejection rates are equal is also used to assess performance. As reported in Table 2, the fused model achieves the lowest EER (0.97%), indicating the best balance between security and convenience, and the adversarial CNN has a substantially lower EER than the plain CNN.

Table 2. Equal Error Rate (EER) by Model

Model	EER (%)
Temporal Classifier	4.10
CNN Classifier	6.91
SSL Classifier	1.06
CNN (Adversarial)	2.56
Fused Model	0.97

4.8 Ablation Study and Model Progression

An ablation study analyzes the contribution of each branch. The temporal model offers fast but weak detection; the CNN model offers better detection but limited robustness; the SSL model generalizes well to new data; and the adversarial CNN is more resistant to attacks. The fusion model retains all of these strengths and delivers the highest overall performance. Removing any branch degrades performance, which demonstrates the effectiveness of multi-branch fusion.

4.9 Interpretability and Feature Analysis

To understand model behavior, interpretability is examined for each branch. The temporal model uses XGBoost feature-importance analysis to reveal the significance of spectral features. The CNN model focuses on spectro-temporal patterns, which can be examined with techniques such as Grad-CAM. The SSL model exploits temporal and contextual information in speech. Overall, the branches capture audio information in terms of spectral features and temporal variation, consistent with prior research on deepfake speech detection. A consolidated comparison against the baseline is given in Table 4.

Table 3. Experimental Results Summary (Binary and Multi-Class)

Model	Binary Acc.	Binary F1 (Real)	Binary F1 (Synth)	Multi Acc.	Multi F1 (Macro)	Multi F1 (Wtd.)	N
CNN	0.9403	0.6071	0.9677	0.8334	0.8263	0.8325	101,921
Temporal (XGBoost)	0.9654	0.8368	0.9807	0.8840	0.8766	0.8834	101,142
SSL Model	0.9851	0.9567	0.9973	0.9858	0.9858	0.9858	101,921
CNN Branch (Adv.)	0.9867	0.9350	0.9926	0.9100	0.9050	0.9090	101,921
Fused Model	0.9930	0.9649	0.9961	0.9420	0.9445	0.9445	101,921

Table 4. Performance Comparison of the Proposed and Baseline Models

Model	Accuracy (%)	Robustness	Feature Representation
Baseline (Spectro-Temporal + CNN Fusion)	94.00	Limited	Spectral + CNN
Proposed CNN (Adversarial)	98.67	Improved	Spectrogram + Adversarial
Proposed SSL Model	98.51	Better generalization	Semantic embeddings
Proposed Fusion Model	99.30	High	Classical + CNN (adv.) + SSL

V. DISCUSSION

The system builds on the spectro-temporal CNN-fusion approach for deepfake speech detection and spoof-system attribution. The baseline primarily uses spectro-temporal features together with CNN representations to achieve high accuracy in detecting deepfake speech and attributing spoofing systems. Although it performs well in standard settings, it does not consider adversarial attacks, does not exploit deep semantic representations from modern self-supervised learning, and limits fusion to spectro-temporal and CNN features.

The proposed system introduces several improvements. First, an SSL branch learns semantic and contextual features from the audio signal. Second,

adversarial training with FGSM is applied to the CNN branch for more robust feature learning. Third, a multi-branch fusion model merges classical features, CNN spectrogram features, and SSL features for a richer representation of the audio. As summarized in Table 3 and Table 4, the proposed system outperforms the baseline across several metrics: the fusion strategy achieves the highest accuracy and the lowest EER, adversarial training substantially improves robustness, and the SSL embeddings improve generalization. In summary, fusion improves detection accuracy and reliability, SSL embeddings improve generalization across spoof types, adversarial training reduces vulnerability to attacks, and the largest gains appear in detection of the bonafide class.

VI. CONCLUSION

This paper presented a deepfake audio detection system that achieves improved accuracy, robustness, and generalization under realistic conditions. The method extends current spectro-temporal and CNN-based techniques with a multi-branch architecture that combines classical feature extraction, CNN-based spectrogram analysis, and self-supervised learning (SSL) embeddings. The classical branch extracts low-level statistical features, the CNN branch models time–frequency structure to capture artifacts in generated speech, and the SSL branch captures semantic characteristics directly from the audio. These signals are combined through score-level fusion, leading to enhanced performance.

A further contribution is adversarial training with the Fast Gradient Sign Method (FGSM). The experiments show that conventional models are susceptible to adversarial attacks, whereas the adversarially hardened CNN remains robust, significantly increasing the security of the detection system. Evaluated on the ASVspoof 2021 LA dataset, the integrated model delivers the best performance, with a binary accuracy of 99.30% and an EER of 0.97%, surpassing the individual models and the baseline. The results confirm that combining SSL embeddings with adversarial training improves both robustness and generalization.

In future work, a multimodal extension based on audio–visual or audio–textual data could further improve accuracy, and the system could be adapted for real-time operation. Adversarial robustness could be strengthened through adaptive adversarial training and certified robustness, and continual learning could help the system handle new spoofing methods. Computational efficiency could be improved through model pruning, quantization, and lightweight architectures, while generalization could be improved by increasing the size and diversity of the training data, including real-world noise and cross-lingual samples.

ACKNOWLEDGMENT

The authors received no specific external funding for this work. Generative AI tools were used to assist with language editing and formatting of this manuscript; all technical content, experiments, analysis, and conclusions are the authors' own work and responsibility, and the authors reviewed and verified the final manuscript.

Conflicts Of Interest: The authors declare no conflict of interest.

REFERENCES

[1] Asuail, C., A. P. Arinomorl, C. T. Atumah, I. F. Kowhor, and D. E. Ogheneochuko, "Hybrid CNN-LSTM architectures for deepfake audio detection using Mel-frequency cepstral coefficients and spectrogram analysis," IEEE Access, vol. 13, pp. 1–13, 2025.

[2] Bohara, R., and A. K. Bairwa, "Detecting deepfake audio using spectrogram-based machine learning approaches," IEEE Access, vol. 13, pp. 1–12, 2025.

[3] Can, Z., and B. Soyhan, "Spectro-temporal-CNN fusion for deepfake speech detection and spoof system attribution," IEEE Access, vol. 13, pp. 1–14, 2025.

[4] Hashmi, A., S. A. Shahzad, C.-W. Lin, Y. Tsao, and H.-M. Wang, "A human-cognition-inspired audio-visual transformer-based ensemble network for video deepfake detection," IEEE Transactions on Information Forensics and Security, vol. 20, pp. 1–16, 2025.

[5] Kaur, N., A. Dixit, and S. Kingra, "A deep learning fusion model leveraging spectral features for audio deepfake detection," Applied Artificial Intelligence, vol. 38, no. 4, pp. 1–14, 2024.

[6] Kwon, H., and S.-H. Nam, "Audio adversarial detection through classification score on speech recognition systems," Computer Speech & Language, vol. 82, pp. 1–12, 2024.
[7] Rabhi, M., S. Bakiras, and R. Di Pietro, "Audio-deepfake detection: Adversarial attacks and countermeasures," Pattern Recognition, vol. 148, pp. 1–15, 2024.
[8] Wang, L., et al., "ERF-BA-TFD+: A multimodal model for audio-visual deepfake detection," IEEE Transactions on Multimedia, vol. 27, pp. 1–15, 2025.
[9] Wani, T. M., and I. Amerini, "Dynamic knowledge condensation with audio selective transformer for audio deepfake detection," Pattern Recognition Letters, vol. 179, pp. 1–10, 2025.