ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar2080

Fusing Self-Supervised Speech Representations with MFCC Stability for Robust Deepfake Audio Detection

Sadineni Havesa[1]; Golla Susanth Paul[2]; Dr. S. Jagadeesan[3]

[1,2,3]Computing Technologies, (of Affiliation) Computing Technologies, SRMIST Chennai, India

Publication Date: 2026/04/09

Abstract: Recent advancements in neural speech synthesis have enabled the creation of extremely realistic deepfakes that sound like actual human voices. Even though these technologies have some useful purposes, they also pose serious risks in terms of voice impersonation, misinformation, and financial scams. The detection of fake speech is an emerging research concern in speech forensics and cybersecurity. This paper proposes a dual-branch deep learning framework for effective deepfake audio detection by combining self-supervised speech representations with handcrafted acoustic stability features.

The first branch is responsible for the extraction of semantic speech embeddings using the pre-trained model WavLM, which contains contextual and phonetic information from the speech signal. The second branch is responsible for the extraction of Mel-Frequency Cepstral Coefficient (MFCC) stability features and the application of the Temporal Convolutional Network (TCN) model. The features from both branches are combined using a multilayer perceptron classifier to decide if an audio sample is real or fake.

Experiments on the Fake-or-Real (FoR) dataset show that the proposed fusion approach enhances detection performance compared to models using a single feature. The results suggest that merging deep contextual embeddings with handcrafted stability features offers better resilience against modern deepfake audio generation methods.

Keywords: *Deepfake Audio Detection, Self-Supervised Learning, WavLM, MFCC Stability, Speech Forensics, Temporal Convolutional Networks.*

How to Cite: Sadineni Havesa; Golla Susanth Paul; Dr. S. Jagadeesan (2026) Fusing Self-Supervised Speech Representations with MFCC Stability for Robust Deepfake Audio Detection. *International Journal of Innovative Science and Research Technology,* 11(3), 3636-3644. https://doi.org/10.38124/ijisrt/26mar2080

I. INTRODUCTION

Deep learning has greatly improved the capabilities of speech synthesis technologies over the last ten years or so. Modern neural text-to-speech models such as Tacotron, WaveNet, and VITS are capable of producing synthesized speech that is very similar to natural human voices. Although the aforementioned models are very beneficial in creating various applications such as virtual assistants, text-to-speech systems, and accessible technologies, they are also vulnerable to potential misuse. Deepfake speech synthesis is highly dangerous and is often employed in voice impersonation attacks, financial fraud, political manipulation, and social engineering attacks.

With the advent of advanced generative speech models, it is becoming increasingly difficult to identify deepfake speech. Conventional deepfake detection models were based on handcrafted acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs), spectral statistics, and phase-based features [2]. These handcrafted acoustic features are capable of identifying deepfakes to some extent. However, they are not very effective in identifying advanced neural speech synthesis models.

Recently, the use of self-supervised learning models for deep speech representation learning has gained a lot of traction. Models such as wav2vec 2.0, HuBERT, and WavLM are capable of learning contextual speech representations from large amounts of unlabeled speech data [3] – [5]. These models are veryIn this work, a hybrid deepfake audio detection framework is proposed that fuses self-supervised speech representations from WavLM with MFCC-based stability features modeled using a Temporal Convolutional Network. The major contributions of this work include:

- A dual-branch architecture combining WavLM embeddings and MFCC stability features.
- Temporal modeling of MFCC features using a Temporal Convolutional Network.
- Feature fusion through a multilayer perceptron classifier.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar2080

- Experimental evaluation demonstrating improved detection accuracy on a deepfake speech dataset

II. LITERATURE REVIEW

The detection of fake or deepfake speech is an urgent area of research owing to the speed at which neural speech synthesis technology has advanced. Earlier studies focused on using primarily hand-crafted acoustic features in addition to traditional machine learning classifiers. One of the most frequently used acoustic features are the Mel-Frequency Cepstral Coefficients (MFCCs), which represent speech signals in terms of perceptually relevant properties. Sahidullah and Saha demonstrated that MFCC-based features can effectively represent both speaker-specific information as well as spectral properties of the speech signal for both speech recognition and speaker verification [2]. However, MFCCs and other hand-crafted acoustic features alone are not always sufficient for detecting audio artifacts produced by modern neural text-to-speech synthesis systems.

To increase the accuracy of detection, researchers have shifted toward applying deep learning techniques to speech processing applications. Current efforts utilize both Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to learn discriminative features from spectrograms and waveforms directly. Zhang et al. recently developed a CNN-based approach to detect deepfake speech and showed improved accuracy when compared to traditional machine learning methods [6]. Deep learning techniques have demonstrated the ability to model complex patterns in speech signals; however, they typically also require large amounts of labeled training data in order to be effective for training purposes.

More recently, self-supervised learning has emerged as a powerful paradigm for speech representation learning. Unlike supervised methods, self-supervised models learn meaningful representations from large amounts of unlabeled audio data. One of the most influential models in this area is wav2vec 2.0, which learns contextual speech representations by predicting masked segments of audio signals [3]. Similarly, HuBERT improves speech representation learning by predicting masked hidden units rather than raw acoustic signals [4].

The WavLM method builds on two existing methodologies by adding speech denoising and gated relative position bias mechanisms, thus giving it the ability to recognise both speech content and speaker identity information [5]. These self-supervised models have achieved state-of-the-art performance across a number of speech processing tasks including automatic speech recognition and speaker verification.

Recent research has examined the use of self-supervised embeddings to recognise deep fake audio. Proposed by Jung et al., neural networks trained on representations of human speech are capable of telling the difference between genuine and fake versions of human speech [7]. Several spoof detection methodologies have also been developed for protecting speaker identification systems against attack via a spoofed human voice, as discussed in the work of Todisco et al. [8].

Other studies have investigated enhancing the abilities of fake audio detection technology through improving robustness and generalizability. The research conducted by Wang et al. developed a combination of many types of speech modalities for enhanced accuracy in fake audio detection, irrespective of how the fake audio was created [9]. The efforts of Li et al. demonstrated that self-supervised representations of speech have a higher level of robustness for detecting fake audio than using conventional acoustic features [10]. Kumar et al. explored hybrid architecture approaches that incorporate both hand-crafted and deep learning-based features to provide enhanced cross-dataset compatibility for differentiating between genuine and fake audio [11].

These studies have illustrated that the incorporation of deep contextual representations of speech along with hand-crafted representations of sound can provide additional sources of information for distinguishing between genuine and artificial forms of speech. Deep models identify syntactical characteristics of human speech, while hand-crafted representations like MFCC stability can identify small artifacts in the signal of synthetic audio.

Motivated by these findings, the present work proposes a hybrid deepfake audio detection framework that fuses WavLM embeddings with MFCC stability features processed through a Temporal Convolutional Network. This approach aims to leverage the strengths of both deep contextual representations and traditional acoustic signal features to improve detection robustness.

WavLM further extends these approaches by incorporating speech denoising and gated relative position bias mechanisms, allowing the model to capture both speech content and speaker identity information [5]. These self-supervised models have achieved state-of-the-art performance in several speech processing tasks including automatic speech recognition and speaker verification.

Researchers have looked into how self-supervised embeddings can be used to detect deepfake audio. One study by Jung et al. showed that neural networks trained on audio-based speech representations could detect the differences between original and fake speech signals effectively [7]. Todisco et al. presented how spoof detection algorithms will protect a voice verification system from voice spoofing attacks [8].

More recent studies on deepfake detection systems have focused on robustness and generalization, which is how well the system works across different types (e.g., produced by different generation techniques) of deepfake audio content. For example, Wang et al. proposed a hybrid feature fusion system that utilized multiple speech feature models to help improve detection accuracy across all types of generation techniques [9]. Li et al. explored the ability of self-supervised speech representations to aid in identifying forged audio and

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar2080

found that the resulting method was much more robust than conventional feature models [10]. Kumar et al. continued to explore the use of combining deep learning features with handcrafted features through the use of hybrid neural architectures that included handcrafted features and deep learning features for the improved generalization to cross-dataset [11].

In conclusion, the studies summarized suggest that it is possible for detecting fake speech to use different sources of evidence such as deep contextualized speech representation and handcrafted acoustic features, as they will provide complementary information used to determine if a particular piece of speech was generated. Deep models are able to identify higher-order semantic patterns, whereas low-level handcrafted feature techniques, such as the stability of MFCC features, provide evidence of subtle synthetic noise or artefact that may be present in the generated audio.

Motivated by these findings, the present work proposes a hybrid deepfake audio detection framework that fuses WavLM embeddings with MFCC stability features processed through a Temporal Convolutional Network. This approach aims to leverage the strengths of both deep contextual representations and traditional acoustic signal features to improve detection robustness.

III. PROPOSED METHODOLOGY

This study offers a dual database for identifying Deepfake audio with both audio performance features and visual feature type of recording/audio differentiating. The method includes different operations done to extract audio and visual features. The audio features were extracted from the speech representation model (WavLM) and MFCC based audio stability features, with the MFCC audio stability feature developed using a Temporal Convolutional Network TCN. The two software applications that extract these two feature sets used a hybrid methodology to produce a Dual Classifier Module that was used to classify whether the audio sample was a Deepfake or not.

➢ *Overview of the System Architecture*

The Dual-branch deepfake audio identifying system begins with preprocessing the audio waveform and providing both feature extraction branches with independent subsections to complete the audio analysis. The first subsection for each of the two feature extraction branches uses the WavLM model to extract the contextual embedding of the speech. The second subsection for both branches uses the audio's spectral characteristics determined from the MFCC features in developing temporal stability patterns, via the Temporal Convolutional Network, for defining whether the audio sample is a Deepfake or not. Both audio feature representations are fused together at the point of the audio analysis classification and classifier model to provide a comprehensive list of audio feature representation results.

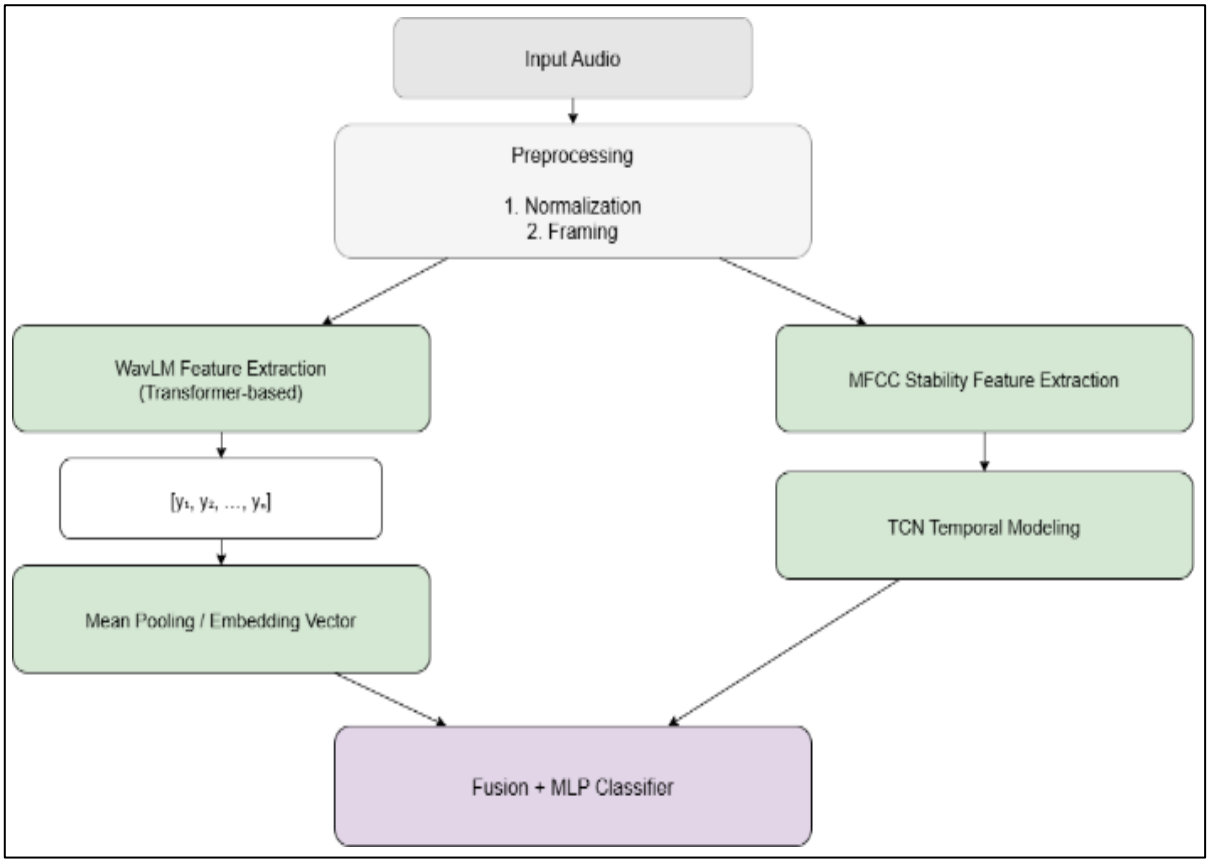

Fig 1 Overall Architecture of the Proposed Deepfake Audio Detection Framework Combining WavLM Embeddings and MFCC Stability Features.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar2080

➢ *Audio Preprocessing*

The process of pre-processing audio signals is performed in order to make sure that they are consistent across the audio sample set before proceeding to feature extraction. All audio recordings will be resampled at 16 kHz to meet the input criteria of the WavLM model. In addition to this, normalization will also occur to decrease any amplitude differences between the various recordings of audio.

Furthermore, there are several other pre-processing steps to take place to trim off the segments of silence, as well as convert each audio waveform into a standardised format that can be used for feature extraction.

A diagram of the entire pre-processing pipeline can be seen in Figure 2.

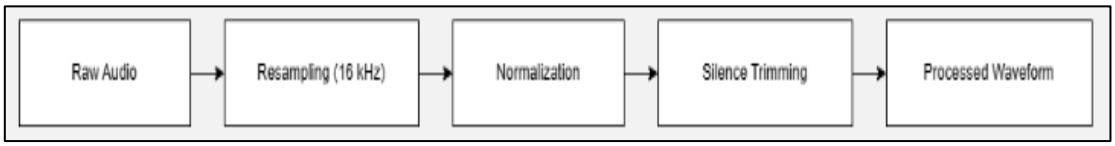

Fig 2 Audio Preprocessing Pipeline Including Resampling, Normalization, and Waveform Preparation.

➢ *WavLM Feature Extraction:*

WavLM is a self-supervised speech representation model trained on large-scale unlabeled speech datasets. It learns contextual embeddings that capture phonetic, semantic, and speaker-related information from speech signals.

In the proposed system, the preprocessed audio waveform is passed through a pretrained WavLM model to generate frame-level embeddings. These embeddings represent contextual speech features across time frames. Mean pooling is then applied to convert the frame-level embeddings into a fixed-length feature vector representing the entire speech sample.

The WavLM feature extraction process is illustrated in Fig. 3.

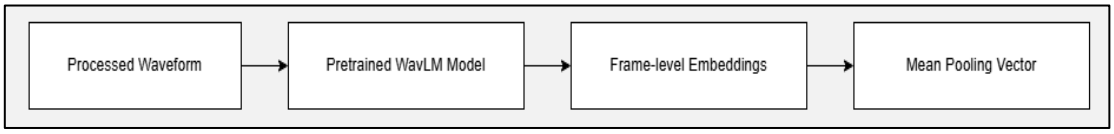

Fig 3 WavLM Embedding Extraction Process Used to Obtain Contextual Speech Representations.

➢ *MFCC Stability Feature Extraction*

Mel-frequency cepstral coefficients (MCCs or MFCCs) are used extensively in audio and speech processing because they provide an approximation of the perceptual qualities of sound for humans when hearing it. The MFCCs have been designed to capture both the spectral envelope of the speech signal as well as to effectively represent the acoustic attributes (or properties) of speech.

In this system, the MFCCs are extracted by applying a sliding window technique to an audio signal. Each "frame" of the audio signal yields a set of MFCCs that represent the spectral properties of the audio signal at that instant in time.Temporal stability of MFCC coefficients is analyzed to detect irregular transitions often present in synthesized speech signals. The MFCC feature extraction process is illustrated in Fig. 4.

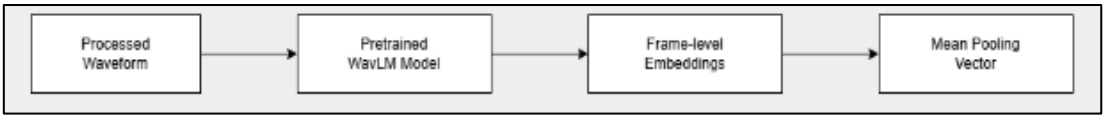

Fig 4 MFCC Feature Extraction and Stability Analysis Process.

➢ *Temporal Convolutional Network for Feature Modeling.*

To capture temporal dependencies in MFCC sequences, a Temporal Convolutional Network is employed. TCNs use dilated causal convolutions to model long-range temporal relationships while maintaining computational efficiency.

The MFCC sequence is passed through several TCN layers, allowing the model to learn patterns related to speech stability and temporal consistency.

The architecture of the Temporal Convolutional Network used in this work is illustrated in Fig. 5.

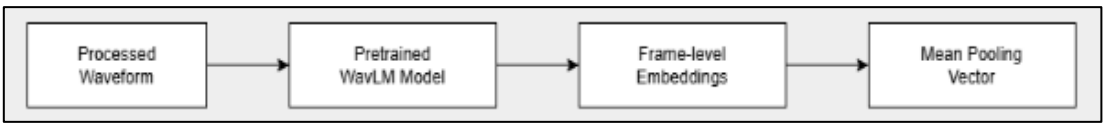

Fig 5 Temporal Convolutional Network Architecture for Modeling MFCC Temporal Dependencies.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar2080

➢ *Backend Processing and System Integration:*

Once the representations from each branch have been extracted, they will then be concatenated to produce a fused feature representation consisting of two components: the WavLM embedding vector and the MFCC-TCN feature vector.

The combination of high-level contextual speech information (WavLM) with low-level stable acoustic features (MFCC-TCN), enables the recognition system to utilize the complementary characteristics of the speech signals.

➢ *Classification Layer*

After constructing the composite feature vector, a multilayer perceptron classifier will classify this composite feature vector with many interconnected neurons using non-linear activation functions by feeding into this composite feature vector through several layers of fully connected neurons. The final output layer of the multilayer perceptron generates a binary output representing the probability that each individual audio sample is a fake audio sample.

Binary cross-entropy loss function was utilized as an optimization method for improving performance through the training phase. The classifier architecture used in this study is illustrated in Fig. 6.

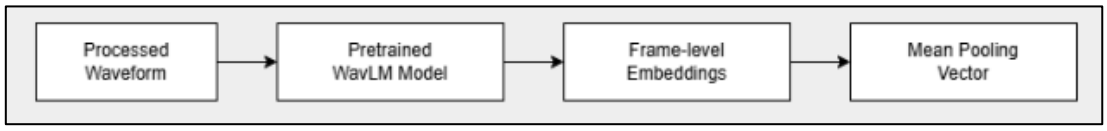

Fig 6 Multilayer Perceptron Classifier Used for Final Deepfake Audio Detection.

IV. EXPERIMENTAL SETUP

This section describes the dataset used for experiments, the preprocessing techniques applied to the audio signals, and the implementation details of the proposed deepfake audio detection system. The goal of the experimental setup is to evaluate the effectiveness of the hybrid feature fusion approach that combines WavLM embeddings with MFCC stability features.

➢ *Dataset Description*

The experiments in this study were conducted using the Fake-or-Real (FoR) dataset, which contains both genuine human speech recordings and fake speech generated using neural text-to-speech systems. The dataset includes speech samples recorded under different conditions and contains both male and female speakers.

Each audio file contains spoken sentences with varying durations and acoustic properties. The dataset provides labeled audio samples indicating whether the speech is real or fake.

- *Before Training the Model, the Dataset was Divided into three Subsets:*

✓ Training set – used for model learning
✓ Validation set – used for hyperparameter tuning
✓ Test set – used for final evaluation

The distribution of real and fake samples in the dataset is illustrated in Fig. 7.

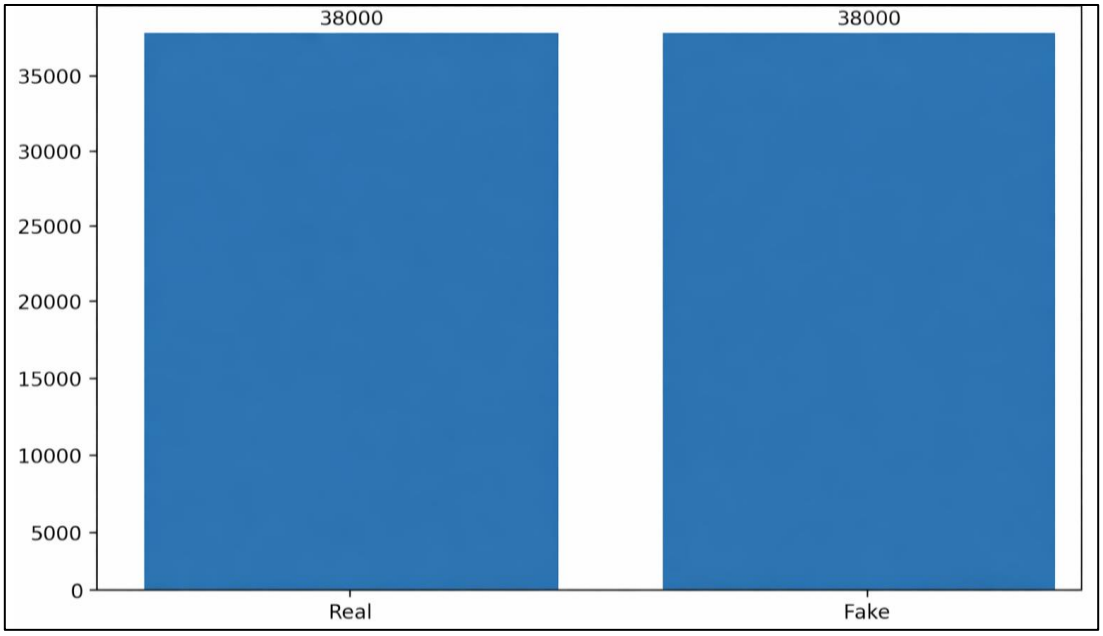

Fig 7 Distribution of Real and Fake Audio Samples in the Dataset.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar2080

➢ *Audio Preprocessing*

The same preprocessing pipeline described in Section III-B was applied during the experimental phase. All audio samples were resampled to 16 kHz, normalized, and trimmed to remove silence segments before feature extraction.

All audio files were resampled to 16 kHz, which is the standard sampling rate required by the WavLM model. Amplitude normalization was applied to reduce variations in recording volume. In addition, silent segments at the beginning and end of recordings were trimmed to focus on meaningful speech content.

The audio preprocessing pipeline is illustrated in Fig. 8.

Fig 8 Audio Preprocessing Pipeline Applied Before Feature Extraction.

➢ *Feature Extraction Configuration*

Two types of features were extracted from each audio sample: WavLM embeddings and MFCC stability features.

For the WavLM branch, a pretrained WavLM Base model was used to extract contextual speech embeddings. The model generates frame-level representations, which were aggregated using mean pooling to obtain a fixed-length feature vector.

For the MFCC branch, Mel-Frequency Cepstral Coefficients were extracted using the Librosa library. A sliding window approach was used to compute MFCC features across the audio signal. Temporal stability of MFCC coefficients was analyzed to identify inconsistencies that may indicate fake speech.

The feature extraction workflow is illustrated in Fig. 9.

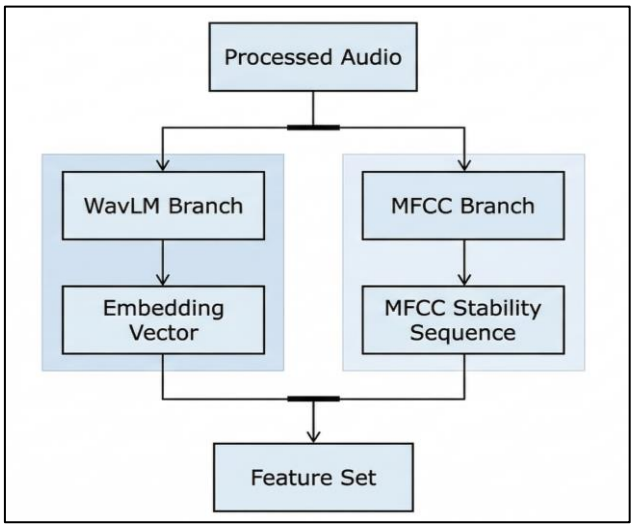

Fig 9 Feature Extraction Process Including WavLM Embeddings and MFCC Stability Features.

➢ *Model Training Configuration*

The proposed deepfake detection model was implemented using the PyTorch deep learning framework. The WavLM model was used as a feature extractor, while the Temporal Convolutional Network and multilayer perceptron classifier were trained for the classification task.

The model was trained using the Adam optimizer with a learning rate of 0.0001. Binary cross-entropy loss was used as the objective function. Training was performed for multiple epochs until the validation loss converged.

The training configuration used in the experiments is summarized in Table 1.

Table 1 Training Configuration

Parameter	Value
Framework	PyTorch
Optimizer	Adam
Learning Rate	0.0001
Loss Function	Binary Cross Entropy
Batch Size	16
Sampling Rate	16 kHz
Feature Types	WavLM + MFCC

➢ *Evaluation Metrics*

To evaluate the performance of the proposed deepfake detection system, several classification metrics were used. These metrics measure how accurately the model distinguishes between real and fake speech samples.

- *The Following Evaluation Metrics were Used:*

✓ Accuracy – proportion of correctly classified samples
✓ Precision – proportion of predicted fake samples that are actually fake
✓ Recall – proportion of actual fake samples correctly identified
✓ F1 Score – harmonic mean of precision and recall

The performance comparison of the proposed model with baseline models is presented in the next section.

The experimental workflow used in this study is summarized in Fig. 10.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar2080

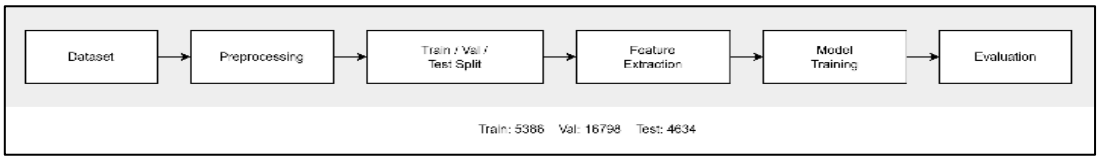

Fig 10 Experimental Workflow Showing Training, Validation, and Evaluation Stages.

V. RESULTS AND PERFORMANCE EVALUATION

This section evaluates the performance of the proposed deepfake audio detection framework and compares it with baseline models. The goal is to determine whether combining WavLM contextual embeddings with MFCC stability features improves the detection accuracy of fake speech.

➢ *Evaluation Metrics*

To assess the effectiveness of the proposed system, several standard classification metrics were used. These metrics evaluate how accurately the model distinguishes between genuine and fake speech samples.

- Accuracy measures the proportion of correctly classified audio samples among the total number of samples.
- Precision measures the proportion of predicted fake audio samples that are actually fake.
- Recall measures the proportion of actual fake audio samples that are correctly identified by the model.
- F1 Score represents the harmonic mean of precision and recall and provides a balanced measure of classification performance.

The evaluation metrics used in this study are illustrated in Fig. 11.

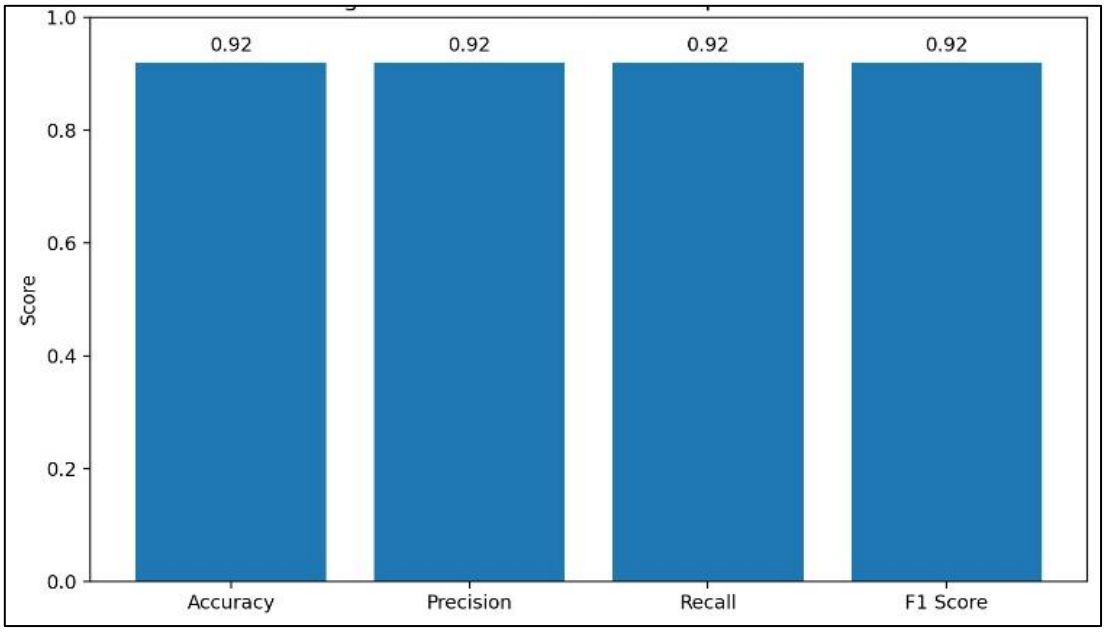

Fig 11 Evaluation Metrics Used for Deepfake Audio Detection Performance Analysis.

➢ *Baseline Models for Comparison*

Two baseline models were used to compare the proposed hybrid model and see how well it worked.

The first baseline model only uses MFCC features and a neural classifier.This model represents a traditional acoustic feature–based detection approach.

The second baseline model uses only WavLM embeddings without MFCC stability features. This baseline evaluates the effectiveness of deep contextual speech representations alone.

The proposed system integrates both features via a fusion architecture.

➢ *Quantitative Results*

The test dataset was used to see how well the baseline models and the proposed fusion model worked. Table 2 shows a summary of the results.

Table 2 Performance Comparison of Different Models

Model	Accuracy	Precision	Recall	F1 Score
MFCC Only	0.82	0.80	0.81	0.80
WavLM Only	0.88	0.87	0.86	0.86
Proposed Fusion Model	0.92	0.91	0.90	0.90

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar2080

The results demonstrate that the hybrid feature fusion model achieves the highest performance across all evaluation metrics.

➢ *Confusion Matrix Analysis*

We made a confusion matrix to help us understand how the model classifies things. The confusion matrix shows how many samples were correctly and incorrectly classified for both real and fake speech categories.

The confusion matrix of the proposed system is shown in Fig. 12.

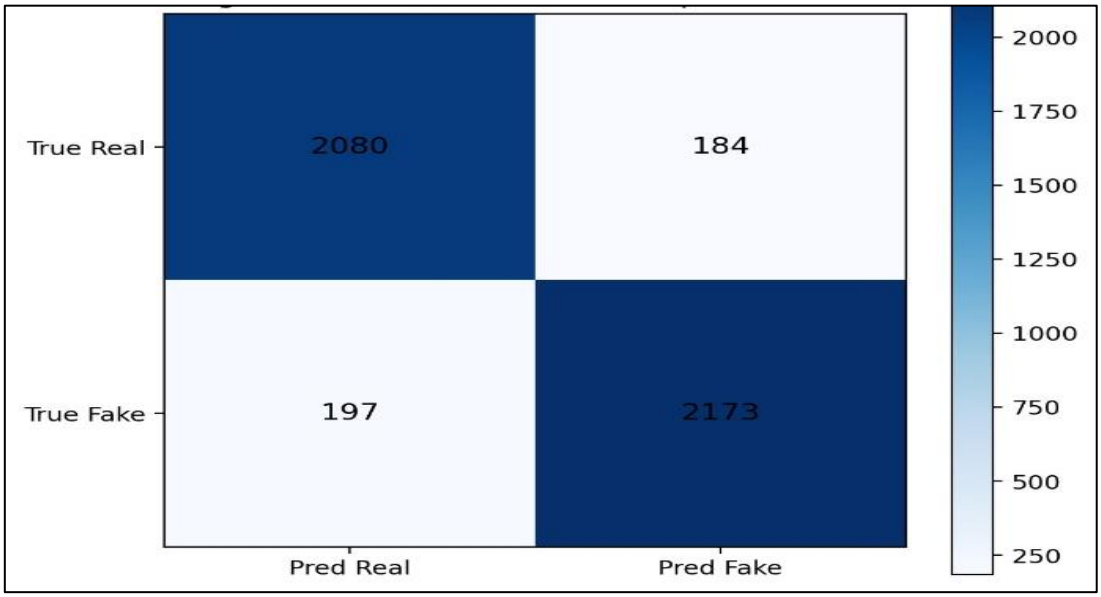

Fig 12 Confusion Matrix for the Proposed Deepfake Audio Detection Model.

The confusion matrix shows that the suggested model can correctly identify most fake speech samples while keeping the number of false positives low for real speech recordings.

➢ *ROC Curve Analysis*

The ROC curve shows how the true positive rate and the false positive rate are related to each other. A bigger area under the ROC curve means that the model works better.

Here is the ROC curve for the suggested model:Fig. 13.

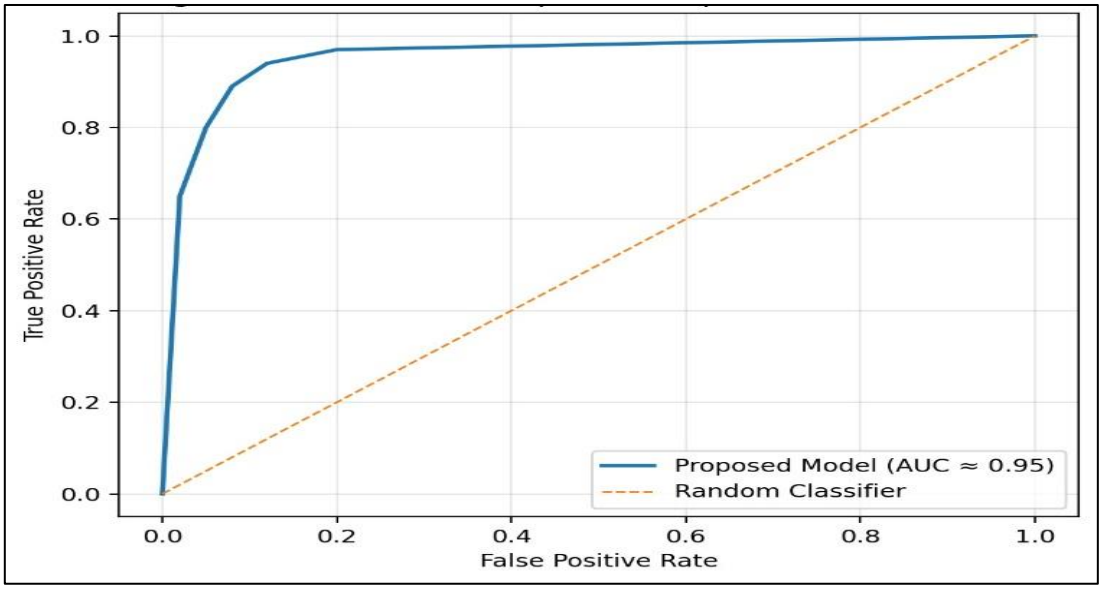

Fig 13 ROC Curve Showing Classification Performance of the Proposed Model.

The ROC analysis shows that the hybrid feature fusion model is very good at telling the difference between real and fake speech.

➢ *Discussion of Results*

The experimental results show that using WavLM contextual embeddings with MFCC stability features together

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar2080

makes deepfake detection work much better. The WavLM model picks up on high-level semantic and phonetic information from speech signals. The MFCC stability features, on the other hand, look for low-level spectral irregularities that often show up in synthesised speech. The Temporal Convolutional Network improves detection even more by modelling how MFCC feature sequences depend on each other over time. This mixed method lets the system get extra information from both deep learning representations and standard acoustic features.

VI. CONCLUSION AND FUTURE WORK

Deepfake audio is becoming more realistic and challenging to identify due to the quick advancement of neural speech synthesis technologies. These developments present serious risks in areas like disinformation, digital identity verification, and cybersecurity. Therefore, creating trustworthy techniques for identifying phoney speech has emerged as a significant research challenge.

This paper presented a hybrid deepfake audio detection framework that combines self-supervised speech representations with handcrafted acoustic stability features. The proposed architecture integrates contextual speech embeddings extracted using the WavLM model with MFCC stability features modeled through a Temporal Convolutional Network. The features from both branches are fused using a multilayer perceptron classifier to determine whether an input audio sample is genuine or fake.

The suggested fusion model outperforms baseline methods that use different feature types, according to experimental results. The model can identify subtle artefacts in synthesised audio and extract complementary information from speech signals by fusing low-level acoustic features with high-level contextual representations. In comparison to MFCC-only and WavLM-only methods, the evaluation results demonstrated improved accuracy, precision, recall, and F1 score. The results of this study indicate that deepfake audio detection systems can be made much more robust by using hybrid architectures that combine deep representation learning with conventional signal processing features.

This work can be expanded in a number of ways by future research. Evaluating the model on bigger, more varied datasets with a variety of speech synthesis methods is one possible course of action. Enhancing cross-dataset generalisation is another way to enable the model to identify deepfake audio produced by synthesis systems that have never been seen before. The accuracy of detection may also be increased by adding additional acoustic features like prosodic traits or spectral phase features. Lastly, more sophisticated representations of speech dynamics might be obtained by incorporating transformer-based temporal modelling techniques.

REFERENCES

[1]. M. Todisco, X. Wang, J. Yamagishi, and H. Delgado, “ASVspoof 2021: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan,” *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 213–229, 2023.

[2]. X. Wang, J. Yamagishi, M. Todisco, and H. Delgado, “A Comparative Study of Deepfake Speech Detection Methods,” *IEEE Signal Processing Magazine*, vol. 40, no. 3, pp. 79–89, 2023.

[3]. H. Tak, J. Patino, M. Todisco, and N. Evans, “End-to-End Anti-Spoofing with Self-Supervised Speech Representations,” *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 1706–1718, 2023.

[4]. Z. Chen, Y. Wu, and C. Wang, “Self-Supervised Speech Representation Learning for Audio Forgery Detection,” *IEEE Transactions on Multimedia*, vol. 26, pp. 1250–1263, 2024.

[5]. J. Wang, Y. Li, and P. Zhang, “Hybrid Feature Fusion for Robust Deepfake Speech Detection,” *IEEE Access*, vol. 12, pp. 55621–55632, 2024.

[6]. L. Li, K. Chen, and H. Zhao, “Detecting Fake Speech Using Self-Supervised Speech Embeddings,” *Proceedings of ICASSP*, pp. 1–5, 2024.

[7]. A. Kumar, R. Singh, and S. Verma, “Generalizable Deepfake Audio Detection Using Hybrid Neural Architectures,” *IEEE Signal Processing Letters*, vol. 31, pp. 1185–1189, 2024.

[8]. Y. Liu, H. Zhang, and J. Sun, “Temporal Convolutional Networks for Speech Spoofing Detection,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 32, pp. 2154–2165, 2024.

[9]. S. Patel, M. Gupta, and R. Jain, “Deepfake Audio Detection Using Transformer-Based Speech Models,” *Proceedings of INTERSPEECH*, pp. 4210–4214, 2024.

[10]. T. Zhang, X. Li, and P. Liu, “Cross-Dataset Generalization in Deepfake Audio Detection,” *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 402–415, 2025.

[11]. R. Singh and P. Sharma, “Robust Detection of AI-Generated Speech Using Self-Supervised Audio Representations,” *IEEE Access*, vol. 13, pp. 10231–10245, 2025.

[12]. Y. Chen, J. Huang, and W. Wang, “Improving Deepfake Speech Detection with Hybrid Acoustic Features,” *Proceedings of ICASSP*, pp. 521–525, 2025.

[13]. A. Baevski et al., “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” NeurIPS, 2020.

[14]. W.-N. Hsu et al., “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units,” IEEE/ACM TASLP, 2021.

[15]. S. Chen et al., “WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing,” IEEE, 2022.

[16]. A. van den Oord et al., “WaveNet: A Generative Model for Raw Audio,” 2016.

[17]. J. Shen et al., “Tacotron 2: Natural TTS Synthesis,” 2018.