International Journal of Safety and Security Engineering
Vol. 15, No. 9, September, 2025, pp. 1827-1835
Journal homepage: http://iieta.org/journals/ijsse

Hybrid CNN-BiLSTM for Deepfake Voice Detection: A Comparative Study

Sidharth Bhroge*, Vijay Mane, Medha Wyawahare, Milind Kamble, Milind Rane, Ruta Sapate, Samruddhi Raut

Department of Electronics and Telecommunication Engineering, Vishwakarma Institute of Technology, Pune 411037, India

Corresponding Author Email: siddharth.bhorge@vit.edu

https://doi.org/10.18280/ijsse.150907

Received: 24 June 2025
Revised: 8 September 2025
Accepted: 18 September 2025
Available online: 30 September 2025

Keywords:
hybrid CNN-BiLSTM, temporal feature extraction, MFCC features, deepfake voice detection

ABSTRACT

The rapid advancement of speech synthesis and voice cloning technologies has raised serious concerns about the misuse of artificial voices in identity fraud, misinformation, and other malicious activities. This research paper proposes a robust system for detecting deepfake voice using time series audio features and deep learning based classifiers. Mel-Frequency Cepstral Coefficients (MFCCs) were extracted from input audio clips to capture critical temporal and spectral characteristics. Several architectures, including Convolutional Neural Network (CNN), Long Short-Term Memory Network (LSTM), Residual Network (ResNet), and a hybrid Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) model, were implemented and evaluated. Unlike existing works that typically rely on either convolutional or recurrent structures alone, our hybrid Convolutional Neural Network-Bidirectional Long Short-Term Memory (CNN-BiLSTM) model combines spatial and sequential learning, enabling it to capture both localized acoustic artifacts and long-range dependencies in speech. It achieved superior performance, with perfect accuracy and minimal validation loss, demonstrating its robustness against subtle manipulations that standalone models often miss. Beyond experimental results, the proposed system has significant practical potential: it can be integrated into voice authentication systems, forensic analysis frameworks, and cybersecurity applications for real-time protection against audio-based fraud. Additionally, the comparative model analysis offers valuable insights for future research, supporting the development of lightweight, generalizable, and deployable solutions for voice integrity assessment.

1. INTRODUCTION

The rapid advancements in Artificial Intelligence (AI) and deep learning have significantly improved the quality of speech synthesis and voice cloning technologies. Text-to-Speech (TTS) systems powered by AI are capable of producing incredibly lifelike human voices, which makes them valuable for uses like accessibility tools, audiobook narration, and virtual assistants. These developments have brought up security issues, though, since bad actors can use them to produce deepfakes that sound like real people, which can result in identity theft, false information, and voice spoofing attacks. Conventional approaches to voice authentication and speaker verification depend on statistical models and manually developed feature extraction methods, which frequently miss minute artifacts in cloned voices. Existing deep learning–based detection approaches are more effective but face several limitations:

(1) They often perform well on a single dataset but lack generalization across different domains,

(2) Many methods are computationally intensive and unsuitable for real-time deployment,

(3) Models that use only convolutional or recurrent architectures tend to capture either local acoustic artifacts or long-term speech dependencies, but not both.

These shortcomings highlight the need for architectures that can achieve high accuracy while maintaining robustness and scalability. The core problem addressed in this study is the reliable detection of deepfake voices in diverse and realistic scenarios, where existing methods struggle with subtle manipulations and real-time applicability.

Our innovation lies in designing and evaluating a hybrid Convolutional Neural Network-Bidirectional Long Short-Term Memory (CNN-BiLSTM) architecture that integrates the strengths of convolutional networks (for spatial feature extraction) and bidirectional recurrent networks (for sequential dependency modeling). This combination enables the detection system to capture fine-grained acoustic distortions as well as long-range temporal patterns in speech. By emphasizing both accuracy and practical deployment potential, the proposed framework advances deepfake audio detection beyond existing approaches.

In order to capture both the natural vocal characteristics and the artificial distortions found in cloned voices, the suggested approach collects important elements from audio signals, such as Mel-Frequency Cepstral Coefficients (MFCCs), Fast Fourier Transform (FFT), and Spectrograms. The model is successful at identifying minute distinctions between actual

and AI-generated sounds because Convolutional Neural Network (CNN) layers examine spatial patterns in these features, and BiLSTM layers gradually learn sequential dependencies.

An actual and cloned voice dataset is preprocessed and separated into training, validation, and testing subsets in order to train and assess the model. The preprocessing stage involves normalization, noise reduction, and feature extraction to ensure that the input data is optimized for learning. The extracted features—MFCCs, FFT, and Spectrograms—are stored in NumPy (.npy) files for efficient access and reduced computational overhead during training. The model training process utilizes Binary Cross-Entropy Loss as the loss function, which is well-suited for binary classification tasks such as distinguishing between real and fake voices. The Adam optimizer is applied to boost training efficiency and convergence speed by dynamically altering learning rates. Learning rate scheduling and early stopping are used to minimize overfitting and enhance model performance. The training is conducted in mini-batches (batch size = 32 or 64) over multiple epochs (typically 10-20 epochs), ensuring that the model learns both low-level and high-level patterns in the audio features. Post-training, the final trained model is saved as deepfake_voice_model.h5 and deployed for real-time inference. During deployment, the system accepts a new audio file as input, processes it through the same feature extraction pipeline, and passes the extracted features to the CNN-BiLSTM model for classification. The model then outputs a probability score, determining whether the audio sample is real or fake. A threshold-based decision is applied to label the input accordingly. To ensure scalability and real-world applicability, the deployed model can be integrated into various voice authentication systems, forensic analysis tools, and cybersecurity frameworks. Additionally, it can be continuously updated with new datasets to improve its ability to detect evolving deepfake techniques.

2. LITERATURE SURVEY

A hybrid Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) model for identifying phony speech recordings was developed and evaluated. Preprocessing methods, including segmentation and normalization, were employed on a balanced dataset that included 5,889 actual and 5,889 false speech samples. Extensive training, validation, and hyperparameter adjustment were performed on the CNN-LSTM model. With a 99.2% accuracy, 99.2% F1-score, 99.4% recall, and 99.0% precision, the experimental findings showed great efficacy [1]. Recent developments in deep learning have greatly enhanced the creation of synthetic voice and speech recognition [2]. Several datasets, including Urban-Sound8K, Conversational, AMI-Corpus, and FakeOrReal, were integrated into a deep learning-based system that was suggested. Four key components comprised the approach:

(1) A CNN-based fake audio detection model with 94% accuracy;
(2) Speaker diarization using Natural Language Processing for text conversion (93% accuracy) and a Recurrent Neural Network (RNN) for speaker labeling (80% accuracy, 0.52 DER);
(3) A speech-denoising module using a Multilayer Perceptron (MLP) and CNN with 93% and 94% accuracy, respectively. As part of the EUCINF (EUropean Cyber and INFormation) project, this study presents a deep learning-based system for detecting deepfake audio.

Three transformation methods were used to turn the raw input audio into spectrograms: Short-time Fourier Transform (STFT), Constant-Q Transform (CQT), and Wavelet Transform (WT), along with auditory-based filters like Mel, Gammatone, Linear Filters (LF), and Discrete Cosine Transform (DCT). Three deep learning approaches were investigated:

(1) Baseline CNN, RNN, and C-RNN models;
(2) Transfer learning using computer vision models like ResNet-18, MobileNet-V3, EfficientNet, and DenseNet-121;
(3) Audio pre-trained models like Whisper, Speechbrain, and Pyannote, with extracted embeddings classified using an MLP.

The best-performing models were fused to improve performance, achieving an Equal Error Rate (EER) of 0.03 [3]. As AI develops quickly, deepfake media production has improved thanks to Generative Adversarial Networks (GANs). Conventional deepfake detection techniques use recurrent networks for temporal analysis and convolutional networks for geographical analysis. SFormer makes use of the Swin Transformer to improve generality across various manipulation techniques and lower computer complexity. Tests on several deepfake datasets, including FF++, DFD, Celeb-DF, DFDC, and Deeper-Forensics, showed excellent performance, with accuracy rates of 100%, 97.81%, 99.1%, 93.67%, and 100%, respectively [4]. Widely utilized in consumer IoT applications, Voice-Driven Devices (VDDs) like Google Home and Amazon Alexa are susceptible to Logical Access (LA) attacks, such as voice conversion and TTS synthesis. A unique Extended Local Ternary Pattern (ELTP) feature descriptor that captures algorithmic artifacts and dynamic vocal tract features in synthetic and converted speech is proposed in this study to address this problem. To improve the ability to distinguish between actual and spoof voices, the ELTP features are merged with Linear Frequency Cepstral Coefficients (LFCC). To distinguish between real and fake signals, a Deep Bidirectional Long Short-Term Memory (DBiLSTM) network is trained using these combined properties [5]. Before deep learning models process audio signals, they are first modified using methods such as spectrograms, MFCCs, and wavelet decomposition. CNNs are used for speaker identification and speech recognition, RNNs are used to record temporal audio patterns, autoencoders are used for feature extraction, transformers are used to analyze temporal and frequency-based features, and hybrid models, which combine deep learning and conventional classifiers, are used. Together, these strategies improve audio classification systems' precision and effectiveness [6]. Detecting AI-generated bogus audio, such as replay assaults, voice conversion, and TTS, is the goal of audio spoofing detection. The lack of generality in traditional machine learning techniques results in inaccurate detection. In order to improve feature extraction from mel-spectrograms, we suggest a deep-layered model that combines VGGish with a Convolutional Block Attention Module (CBAM). This architecture captures spatial and channel-based correlations to properly classify audio as real or fake. Our model's efficiency in identifying audio deepfakes was demonstrated by its EER, which was 0.52% for Physical Access (PA) attacks and 0.07% for LA

assaults when tested on the ASVspoof 2019 dataset [7]. Deepfake content, including highly realistic images, audio, and videos created using AI, poses serious threats to national security, democracy, and personal privacy. The paper categorizes deepfake generation methods into face swapping and facial reenactment, while detection approaches primarily rely on feature-based and machine learning techniques. However, deepfake detection still faces challenges such as the continuous improvement of deepfake generation techniques, the scarcity of high-quality datasets, and the absence of standardized benchmarks [8]. Support Vector Machines (SVMs), Decision Trees (DTs), CNNs, Siamese CNNs, Deep Neural Networks (DNNs), and hybrid CNN-RNN architectures are examples of detection techniques, along with statistical analysis and media consistency checks. According to performance evaluation across research, DT had the lowest accuracy at 73.33%, while SVM had the greatest at 99%. Siamese CNNs demonstrated a 55% improvement in the tandem detection cost function (t-DCF), while Deep-Sonar's EER ranged from 2% to 12.24% [9]. With uses ranging from helping with speech impairments to facilitating financial fraud and fake news, fake voices have emerged as a significant problem in social media, cybersecurity, and forensics. A deep learning and machine learning-based approach to phony speech detection is suggested as a countermeasure. To identify whether a voice is synthetic or real, MFCCs are taken out as features and categorized using several machine learning and deep learning models. The efficiency of this method in differentiating between real and phony voices is confirmed by experimental results [10]. Conventional detection techniques concentrate on intra-modal artifacts, whereas multi-modal analysis is necessary for real-world deepfakes. In order to detect deepfakes, this research suggests AVoiD-DF, an Audio-Visual Joint Learning technique that takes advantage of audio-visual discrepancies. The framework consists of a Cross-Modal Classifier for identifying discrepancies, a Multi-Modal Joint-Decoder for fusion, and a Temporal-Spatial Encoder for feature extraction. A new benchmark dataset, DefakeAVMiT, spanning a variety of forgery strategies across modalities, is presented to improve evaluation.

DefakeAVMiT, FakeAVCeleb, and DFDC experimental results show that AVoiD-DF outperforms state-of-the-art techniques and exhibits strong generalization across various forgery strategies [11]. This study presents BTS-E, a framework that assesses the relationship between breathing, talking (speech), and silence sounds in an audio clip in order to improve deepfake detection. The efficacy of this method is confirmed by extensive tests on the ASVspoof 2019 and 2021 datasets, which show a 46% improvement in classifier performance [12]. ASVspoof Challenges have prompted studies into detecting techniques due to the increasing dangers posed by deepfake audio. However, the speech deepfake (DF) subset, which includes a variety of spoof audio sources, continues to be a challenge for state-of-the-art models. In order to improve generality and resilience, this study suggests a unique detection architecture. To enhance representation, the method combines hand-crafted features with learnt embeddings [13]. The challenge of synthetic voice and spliced audio spoofing is addressed by improving existing countermeasures. Traditional methods treat detection as a binary classification problem (bonafide vs. spoof), but this work extends Res2Net with a Conformer block to better capture local acoustic patterns. The proposed SE-Res2Net-Conformer architecture significantly improves detection performance for LA spoofing, according to experimental results on the ASVspoof 2019 dataset. The paper also suggests a novel approach to audio splicing detection, which shifts the focus from identifying entire spliced segments to detecting their precise boundaries [14]. The paper suggests identifying artificially manufactured sounds by utilizing CNNs with Mel spectrograms. To find the best CNN architecture for language-independent detection, supervised experiments were carried out utilizing voice samples from several datasets. The greatest accuracy ratings were 98% for WaveFake, 94% for Automatic Speaker Verification (ASV), and 99% for the FoR dataset. The model's accuracy was 98% for FoR, 92% for ASV, and 96% for WaveFake when it was trained on all datasets at once and tested on separate datasets [15]. Recent advancements in speech synthesis and voice cloning have made it increasingly difficult to distinguish between human and machine-generated speech. As a result, audio deepfake detection has emerged as a critical area of research in audio forensics and AI safety. Wu et al. [16] provided a comprehensive survey of spoofing attacks and countermeasures for ASV systems. Their work highlighted the evolution of synthetic speech generation and the growing need for reliable detection frameworks. To tackle this, large-scale pretrained models such as Pretrained Audio Neural Networks (PANNs) have been introduced by Kong et al. [17], significantly improving performance in audio event recognition tasks by transferring knowledge across audio domains.

Wani and Amerini [18] developed an Adaptive Spectro-Temporal Diffusion Transformer (ASTDT), merging transformer-based attention with diffusion-based modeling for interpretable and robust detection under noisy conditions.

Liu et al. [19] presented findings from the ASVspoof 2021 challenge, outlining benchmarks for synthetic and replayed audio detection in real-world settings. Tahaoglu et al. [20] investigated spectral feature-based deepfake detection, demonstrating how high-frequency anomalies in generated audio can serve as reliable indicators of manipulation. Rabhi et al. [21] conducted one of the first comprehensive studies on adversarial vulnerabilities in deepfake detectors, revealing how imperceptible perturbations could deceive CNN-based classifiers and proposing lightweight adversarial training as a countermeasure. Sharafudeen et al. [22] contributed a feature-fusion framework blending spectral and Bag-of-Audio-Blocks (BoAB) features with BiLSTM networks, achieving superior robustness against spoofing attacks.

Yi et al. [23] provided updated literature reviews summarizing recent trends in audio deepfake generation and detection, emphasizing the importance of explainable AI in forensics. Li et al. [24] proposed SafeEar, a content privacy-preserving model capable of detecting deepfakes while masking semantic content, ensuring ethical compliance and user confidentiality. Wang et al. [25] developed AVT2-DWF, a multimodal detection architecture employing dynamic weight fusion to adaptively integrate visual and auditory modalities for improved deepfake detection in unconstrained environments.

Katamneni and Rattani [26] introduced MIS-AVoiDD, a model disentangling modality-specific and shared audiovisual features, leading to higher cross-modal interpretability. Yang et al. [27] contributed to forensic deepfake analysis by using segmental acoustic features linked to articulatory phonetics, enhancing interpretability in court-admissible forensic contexts.

Recent advancements emphasize end-to-end and

perceptually grounded modeling. Di Pierno et al. [28] introduced a RawNet-based architecture trained directly on raw waveforms, bypassing handcrafted feature extraction and demonstrating superior cross-dataset generalization. El Kheir et al. [29] proposed Two Views, One Truth, a fusion model combining self-supervised representations with spectral handcrafted features (MFCC, LFCC, CQCC) via cross-attention, achieving state-of-the-art robustness against unseen manipulations. Complementing these, Uddin et al. [30] conducted a systematic adversarial benchmarking study across multiple detectors, revealing widespread susceptibility to gradient-based attacks and highlighting the need for adversarially trained, perceptually aware defenses. Their findings confirm that integrating adversarial training with perceptual acoustic fingerprints can create generalizable, dual-layer defenses against evolving deepfake generation methods.

3. RESEARCH GAP

The literature evaluation identifies many research needs in the realm of deepfake audio detection and production. For starters, there is a lack of generalization across datasets and manipulation approaches, with models frequently performing well on specific datasets but not universally. Second, many techniques are not tuned for real-time detection, which is critical for applications such as voice assistants and security systems. Furthermore, current models struggle to recognize subtle and complicated deepfakes, needing more advanced detection approaches. Integration of multi-modal detection techniques, which include auditory and visual signals, is likewise underexplored, despite its potential to improve accuracy. The lack of high-quality and diverse datasets impedes detection model training and evaluation, emphasizing the necessity for larger datasets. The ethical and legal implications of deepfake technologies remain unexplored, necessitating additional research to build norms and policies.

Traditional feature extraction methods may not adequately capture the intricacies of deepfake audio, indicating the need for new strategies. Existing countermeasures, such as binary classification models, require development in order to successfully meet the complexities of spoofing attacks. Addressing these deficiencies may result in more robust, efficient, and generalizable deepfake audio detection systems, improving the security and dependability of audio-based applications.

3.1 Problem statement

The objectives of our research are to:

- Create and evaluate a CNN-LSTM model for identifying phony speech recordings.
- Implement effective preprocessing methods such as segmentation and normalization on a balanced dataset to improve model performance.
- Explore and compare different deep learning approaches, including baseline CNN, RNN, and C-RNN models, transfer learning using computer vision models, and audio pre-trained models.
- Optimize the detection model for real-time applications, ensuring efficient and fast processing without compromising accuracy.

4. METHODOLOGY

4.1 Dataset

This study employs a publicly available deepfake voice detection dataset, comprising raw audio samples organized into two distinct classes: REAL and FAKE, located in the "AUDIO" directory. Each audio file is named to reflect the speaker's identity and whether it is a real or synthetic (deepfake) voice.

For structured analysis, we utilize the DATASET-balanced.csv file, which contains pre-extracted features from one-second audio segments. To ensure class parity, the data was balanced through random sampling, achieving an equal distribution between genuine and deepfake samples. This balanced dataset serves as the foundation for our feature extraction and model training pipeline.

4.1.1 Data preprocessing

The preprocessing of audio data is a crucial step in ensuring model robustness for deepfake voice detection.

The procedure consists of the following stages:

i. Noise Reduction and Normalization: Raw audio files undergo denoising to eliminate background interference, followed by amplitude normalization to standardize loudness across samples.
ii. Segmentation: Audio files are segmented into fixed one-second windows to facilitate frame-wise analysis. Corrupted or silent segments are excluded.
iii. Feature Transformation: Each segment is converted into an efficient numerical representation using audio feature extraction techniques (e.g., MFCCs, spectral features) suitable for input into machine learning models.
iv. Label Encoding: Preprocessed samples are then labeled and organized into REAL and FAKE categories for supervised learning.

This pipeline ensures the data is clean, standardized, and semantically meaningful, preparing it for effective training and evaluation of detection models.

4.1.2 SMOTE-based balancing

The Class imbalance is a common challenge in real-world datasets, where genuine audio typically outnumbers synthetic samples. To address this, we employed the Synthetic Minority Over-sampling Technique (SMOTE). Unlike simple oversampling, which duplicates minority class samples and risks overfitting, SMOTE generates new synthetic samples by interpolating between existing ones in feature space. This approach increases diversity in the training data, allowing the model to learn more generalized decision boundaries. SMOTE was chosen over under sampling (which discards valuable real samples) and random oversampling (which may create bias) because it improves classifier sensitivity to minority classes while maintaining dataset richness. This makes SMOTE particularly suitable for deepfake detection, where the ability to detect underrepresented fake voices is critical for robust performance.

Table 1 shows the structure of the dataset split in the referenced deepfake image detection framework; a similar division has been applied to our audio-based dataset. From a total of 5,000 real and 5,000 fake voice samples, data has been split into 70% for training, 20% for validation, and 10% for testing.

Table 1. Dataset split for real and fake voice samples

Category	Train	Validation	Test
Real	3500	1000	500
Fake	3500	1000	500

4.2 MFCC feature extraction

To improve the detection of deepfake voices, we utilize MFCCs, a widely adopted technique in speech and audio processing. MFCCs are effective because they capture the essential spectral characteristics of an audio signal in a manner that closely mimics human auditory perception. The extraction of MFCCs involves transforming raw audio data into a series of coefficients that represent the power spectrum of short, overlapping windows of audio. This process significantly reduces the dimensionality of the data while preserving key acoustic features, making it easier to differentiate between authentic and synthetic voices.

The MFCC extraction technique is particularly advantageous for deepfake detection due to its sensitivity to subtle variations in vocal characteristics. By emphasizing these coefficients, we can enhance the model's ability to detect even the most minor manipulations in voice recordings. This, in turn, contributes to improved detection accuracy and reliability, enabling the model to better identify discrepancies between real and artificially generated voices.

4.3 Model architecture

To determine the most effective approach for deepfake voice detection, several neural network architectures were designed and evaluated, each with unique strengths in capturing different aspects of the audio features. The following models were implemented:

4.3.1 CNN model

The CNN model is effective for capturing localized spatial patterns in MFCC features, such as frequency variations and short-term distortions. These subtle artifacts often reveal synthetic manipulations that are not perceptible to the human It employed 1D convolutional layers with ReLU activations, allowing the model to learn spatial hierarchies in the data. Max pooling layers were incorporated to reduce the dimensionality of the feature maps, while batch normalization layers helped stabilize training and improve convergence. Dropout was used as a regularization technique to prevent overfitting by randomly deactivating a fraction of neurons during training.

4.3.2 LSTM model

The LSTM model was designed to capture long-range temporal dependencies present in speech patterns. LSTM cells are particularly effective for learning sequential relationships in data, such as the rhythm and timing of speech, which can often be altered in synthetic voice recordings. By utilizing one or more LSTM layers, this model was able to model the dynamic temporal characteristics of the voice, making it adept at detecting subtle irregularities in speech sequences that might indicate deepfake manipulation.

4.3.3 ResNet model

A 1D variant of the ResNet was implemented to facilitate the learning of deeper network architectures. ResNet employs residual connections that allow gradients to flow more efficiently through the network, mitigating the problem of vanishing gradients in deep networks. This design not only improved the model's ability to learn complex representations but also enabled the training of deeper networks, which in turn enhanced its ability to classify deepfake voices with greater accuracy.

4.3.4 CNN-BiLSTM hybrid model

The CNN-BiLSTM hybrid model was developed to combine the strengths of both convolutional and recurrent neural networks. CNN layers extract robust local acoustic features, while BiLSTM layers capture contextual temporal dependencies in both forward and backward directions. This combination allows the model to detect fine-grained spectral anomalies as well as long-term sequence inconsistencies—two critical indicators of synthetic audio. Unlike standalone CNN or LSTM models, the hybrid structure provides a balanced and comprehensive approach, leading to superior performance and stronger generalization. Making it well-suited for capturing both local patterns and long-range temporal dependencies in the audio, enhancing its effectiveness in detecting deepfake voices.

Model Training and Optimization

All models were trained using Binary Cross-Entropy as the loss function, which is commonly used for binary classification tasks such as deepfake voice detection. The Adam optimizer was employed for gradient descent optimization, with a learning rate scheduler to adjust the learning rate during training, ensuring efficient convergence. To prevent overfitting, early stopping was applied, halting the training process if the validation performance stopped improving for a predefined number of epochs. This approach ensured that the models were able to generalize well to unseen data and maintained robustness against overfitting during training.

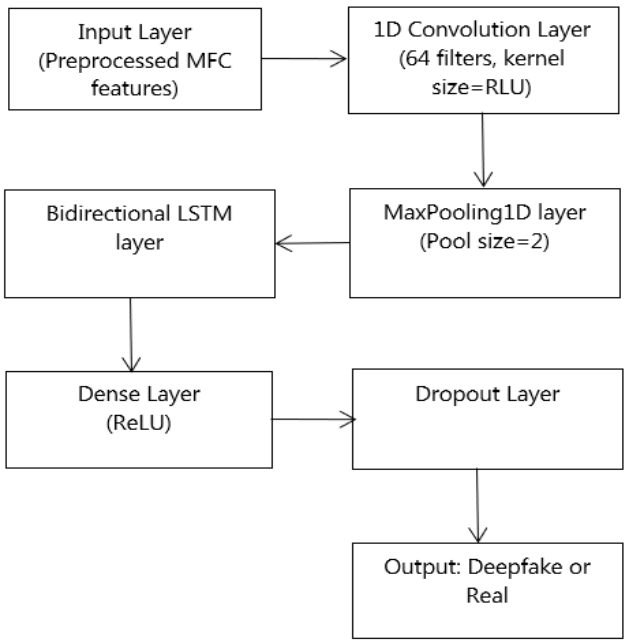

Figure 1. CNN-BiLSTM-based deepfake voice detection architecture

As shown in Figure 1, the proposed deepfake detection model employs a hybrid CNN and BiLSTM architecture. The use of MFC features as input enables effective capture of

spectral properties, while the combination of convolutional and recurrent layers enhances both spatial and temporal feature learning.

Block diagram of the proposed deepfake voice detection model using a hybrid CNN-BiLSTM architecture. The model processes preprocessed Mel-Frequency Cepstral (MFC) features through convolutional and pooling layers for feature extraction, followed by a bidirectional LSTM to capture temporal dependencies. A dense layer with dropout helps in robust classification, and a sigmoid output predicts the probability of a voice being real or a deepfake.

4.4 Deepfake detection

The following algorithm presents the proposed deepfake voice detection pipeline, covering all essential stages from preprocessing to classification and model evaluation:

Algorithm: Deepfake Voice Detection Using MFCC and Deep Learning

Input: Labeled audio samples from the Kaggle Deepfake Voice Dataset

Output: Classification label (*Real or Fake)

Step 1: Dataset Preparation

Acquire and clean audio data.

Normalize, segment (1-sec windows), and label as REAL/FAKE.

Step 2: Feature Extraction (MFCC)

Extract Mel-Frequency Cepstral Coefficients from each audio segment.

Store MFCC matrices as model inputs.

Step 3: Data Balancing

Apply SMOTE to oversample the minority class for balanced learning.

Step 4: Model Construction

Build and compare models:

• CNN – spatial feature extraction

• LSTM/BiLSTM – temporal modeling

• Residual Network – deeper learning with skip connections

• Hybrid CNN+BiLSTM – combined spatial-temporal modeling

Step 5: Training & Validation

Train on 70% data, validate on 20%.

Use early stopping and dropout to avoid overfitting.

Step 6: Evaluation

Assess each model using Accuracy, Precision, Recall, F1-Score, and AUC.

Visualize with a confusion matrix and training curves.

Step 7: Classification & Selection

Choose the best model (e.g., Hybrid CNN+BiLSTM) for final predictions.

Classify unseen audio as Real or Fake.

The model is trained on different Deep Learning Models. After training, the model has been exported and tested on the test dataset, which consists of 5,000 voices containing both real and fake. Then the accuracy of the model has been calculated.

Our approach to deepfake detection follows a structured and well-thought-out flow. It all starts with the data collection of both real and fake voices that form the basis of our system. To make this data useful, we take a step called preprocessing, where we divide it into three key parts: the training set, the validation set, and the test set. We distribute them in a balanced way, with 70% for training, 20% for validation, and 10% for testing. CNN+BiLSTM is an efective technique that we utilize to train the model using the training dataset. We used exported models for deepfake detection in the testing set. This indicated the true effectiveness of our deepfake detecting algorithm. It served as performance evaluation of the system, demonstrating its dependability and efficiency in exposing misleading material across a range of contexts.

The hybrid CNN+BiLSTM model was trained using the following configuration:

- Learning Rate: 0.0001
- Activation Function: ReLU
- Optimizer: Adam
- Batch Size: 64
- Epochs: 10

The Adam optimizer, known for combining the advantages of AdaGrad and RMSProp, was selected for efficient gradient descent. A low learning rate of 0.0001 ensured gradual convergence and stability during training. The ReLU activation function was used in convolutional and dense layers for improved non-linearity and convergence. The model was trained with a batch size of 64, which balances performance with GPU memory efficiency. Training was conducted over 10 epochs to prevent overfitting while ensuring adequate learning.

5. RESULT AND DISCUSSION

To evaluate the performance of various deep learning architectures for deepfake voice detection, four different models were implemented and trained: LSTM-only, CNN-only, ResNet, and a hybrid CNN+BiLSTM model. Each model was trained and validated on the same dataset comprising extracted MFCC features. Table 2 presents a comprehensive comparison of each model's final training and validation metrics.

The CNN+BiLSTM hybrid model outperformed all other architectures, achieving perfect training and validation accuracy (100%) with minimal loss, indicating excellent generalization and effective learning of both spatial features (via convolutional layers) and temporal dependencies (via BiLSTM layers). Its low validation loss of 0.0763 suggests a well-regularized model capable of robust prediction on unseen data.

While CNN-only and LSTM-only models performed reasonably well—each achieving approximately 90% training accuracy and 84% validation accuracy—their higher validation losses (0.2615 and 0.4267, respectively) indicate relatively less robustness and increased risk of overfitting compared to the hybrid model. These models were able to capture either spatial or sequential patterns individually, but not both.

The ResNet model, despite its success in image-related tasks, performed poorly on 1D MFCC features. Its 2D convolutional design is inherently optimized for spatial image structures rather than sequential audio signals. As a result, it failed to learn discriminative temporal patterns, reflected in its high training and validation losses. Additionally, ResNet's deeper architecture made it prone to overfitting in this domain, given the limited variability of MFCC-based representations compared to images. It yields both high training loss (3.6122) and validation loss (5.4762), suggesting either overfitting or a

mismatch in model architecture and data structure. This highlights the need to tailor network designs to the nature of the input—ResNet's 2D convolutional blocks are not optimized for sequential 1D MFCC data.

Table 2. Comparison of detection accuracies of CNN models tested on various GANs

Model	Epochs	Train Accuracy	Val Accuracy	Train Loss	Val Loss
LSTM	20	89.03%	84.62%	0.3022	0.42
CNN	20	90.34%	84.62%	0.2676	0.2615
ResNet	20	85.64%	84.62%	3.6122	5.4762
CNN+BiLSTM	60	100.00%	100.00%	0.0095	0.0763

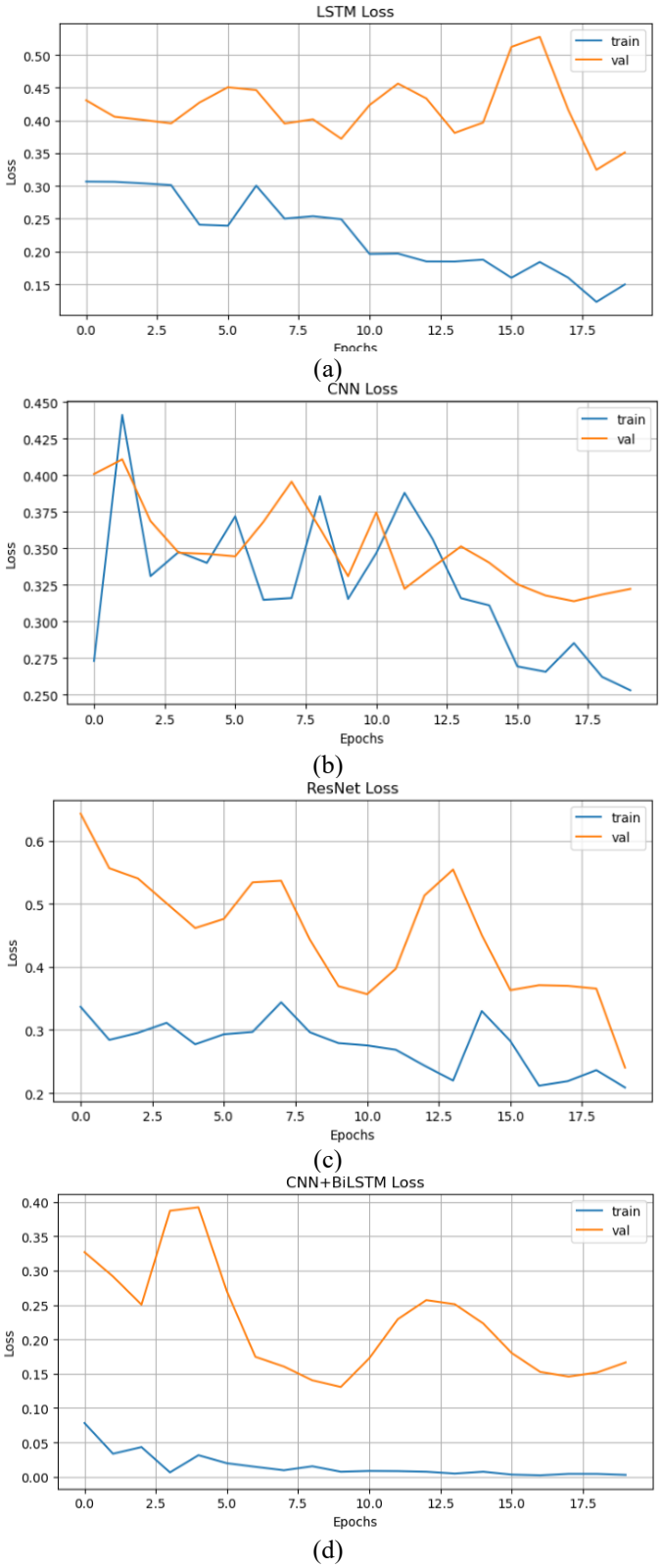

Figure 2. Loss curve for each model

Training and Validation Loss Curves

To further understand the training dynamics, Figure 2 illustrates the loss curves across epochs for each model. The CNN+BiLSTM model demonstrates rapid convergence and minimal divergence between training and validation loss, indicating strong generalization. In contrast, the ResNet model exhibits volatile loss patterns and a lack of convergence. The LSTM and CNN models show moderate overfitting, with training and validation loss diverging after several epochs.

The Long Short-Term Memory (LSTM) model's loss curves shown in Figure 2(a) indicate a slight but steady decrease in both training (blue) and validation (orange) loss over the course of training. The validation loss remains relatively stable around 0.4-0.45 with minor fluctuations, suggesting consistent validation performance. The low difference between training and validation loss implies minimal overfitting, although the model's convergence appears gradual.

This plot is presented in Figure 2(b), the loss trajectories for the CNN model across 20 training epochs. The training loss (blue) shows a general decreasing trend with fluctuations, indicating progressive learning. The validation loss (orange) initially peaks early in training, then gradually decreases, demonstrating the model's improved generalization over epochs. Notably, the training loss consistently remains lower than validation loss, suggesting effective learning with some degree of overfitting in later epochs.

The ResNet shown in Figure 2(c) demonstrates effective training, with both training (blue) and validation (orange) losses decreasing substantially across epochs. The validation loss exhibits some fluctuations but overall shows a decreasing trend from approximately 0.6 to 0.4, indicating robust learning and generalization. The consistently lower training loss reflects good model fit and training stability.

This combined model, shown in Figure 2(d), integrates CNN and Bidirectional LSTM components, with training loss (blue) markedly lower than validation loss (orange). Both losses decrease over epochs, with training loss approaching near zero (~0.05), indicative of efficient training. Validation loss stabilizes around 0.15–0.2, reflecting reasonable model generalization. The divergence between training and validation loss highlights typical overfitting behavior, mitigated through model design or regularization strategies.

The superior performance of the hybrid model reaffirms the effectiveness of combining convolutional layers for feature extraction with bidirectional LSTMs for capturing time-dependent speech patterns, aligning with findings from recent literature in speech-based deepfake detection [3, 4].

6. CONCLUSIONS

This study presented a comparative evaluation of four deep learning models—CNN, LSTM, ResNet, and a hybrid CNN+BiLSTM—for the task of detecting deepfake audio

using MFCC feature representations. The experimental results demonstrate that the CNN+BiLSTM hybrid model achieved the best performance, with 100% accuracy on both training and validation datasets and significantly lower loss values. This indicates its strong ability to capture both local spatial and long-term temporal dependencies in sequential audio data.

In contrast, standalone CNN and LSTM models attained reasonable classification performance (~90% training accuracy), but exhibited higher validation loss, suggesting comparatively weaker generalization. The ResNet architecture, although effective in image classification tasks, did not perform well with 1D MFCC features, resulting in higher training and validation loss. This highlights the importance of aligning model architecture with the nature of the input data.

The proposed hybrid model can be integrated into real-world systems such as voice authentication platforms, call center fraud detection, forensic analysis, and cybersecurity frameworks. Its efficiency and ability to process audio in real time make it particularly suitable for securing VDDs and digital identity systems

Overall, the findings suggest that hybrid architectures combining convolutional and recurrent layers are more effective for audio-based deepfake detection. Future work will explore: a) Using of lightweight architectures and deployment strategies for real-time applications; b) Developing larger and more diverse datasets covering multiple languages, accents, and environmental conditions; c) Developing larger and more diverse datasets covering multiple languages, accents, and environmental conditions; d) Extending detection to multimodal approaches by combining audio with video or textual data.

REFERENCES

[1] Oyucu, S., Çelımlİ, D.B.Ü., Aksöz, A. (2024). Fake voice detection: A hybrid CNN-LSTM based deep learning approach. In 2024 8th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Turkiye, pp. 1-6. https://doi.org/10.1109/ISMSIT63511.2024.10757293

[2] Wijethunga, R.L.M.A.P.C., Matheesha, D.M.K., Al Noman, A., De Silva, K.H.V.T.A., Tissera, M., Rupasinghe, L. (2020). Deepfake audio detection: A deep learning based solution for group conversations. In 2020 2nd International Conference on Advancements in Computing (ICAC), Malabe, Sri Lanka, pp. 192-197. https://doi.org/10.1109/ICAC51239.2020.9357161

[3] Pham, L., Lam, P., Nguyen, T., Nguyen, H., Schindler, A. (2024). Deepfake audio detection using spectrogram-based feature and ensemble of deep learning models. In 2024 IEEE 5th International Symposium on the Internet of Sounds (IS2), Erlangen, Germany, pp. 1-5. https://doi.org/10.1109/IS262782.2024.10704095

[4] Kingra, S., Aggarwal, N., Kaur, N. (2024). SFormer: An end-to-end spatio-temporal transformer architecture for deepfake detection. Forensic Science International: Digital Investigation, 51: 301817. https://doi.org/10.1016/j.fsidi.2024.301817

[5] Arif, T., Javed, A., Alhameed, M., Jeribi, F., Tahir, A. (2021). Voice spoofing countermeasure for logical access attacks detection. IEEE Access, 9: 162857-162868. https://doi.org/10.1109/ACCESS.2021.3133134

[6] Zaman, K., Sah, M., Direkoglu, C., Unoki, M. (2023). A survey of audio classification using deep learning. IEEE Access, 11: 106620-106649. https://doi.org/10.1109/ACCESS.2023.3318015

[7] Kanwal, T., Mahum, R., AlSalman, A.M., Sharaf, M., Hassan, H. (2024). Fake speech detection using VGGish with attention block. EURASIP Journal on Audio, Speech, and Music Processing, 2024(1): 35. https://doi.org/10.1186/s13636-024-00348-4

[8] Zhang, T. (2022). Deepfake generation and detection, a survey. Multimedia Tools and Applications, 81(5): 6259-6276. https://doi.org/10.1007/s11042-021-11733-y

[9] Shaaban, O.A., Yildirim, R., Alguttar, A.A. (2023). Audio deepfake approaches. IEEE Access, 11: 132652-132682. https://doi.org/10.1109/ACCESS.2023.3333866

[10] Kilinc, H.H., Kaledibi, F. (2023). Audio deepfake detection by using machine and deep learning. In 2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM), Istanbul, Turkiye, pp. 1-5. https://doi.org/10.1109/WINCOM59760.2023.10323004

[11] Yang, W., Zhou, X., Chen, Z., Guo, B., et al. (2023). Avoid-DF: Audio-visual joint learning for detecting deepfake. IEEE Transactions on Information Forensics and Security, 18: 2015-2029. https://doi.org/10.1109/TIFS.2023.3262148

[12] Doan, T.P., Nguyen-Vu, L., Jung, S., Hong, K. (2023). BTS-E: Audio deepfake detection using breathing-talking-silence encoder. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, pp. 1-5. https://doi.org/10.1109/ICASSP49357.2023.10095927

[13] Li, M., Ahmadiadli, Y., Zhang, X.P. (2023). Robust deepfake audio detection via Bi-level optimization. In 2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP), Poitiers, France, pp. 1-6. https://doi.org/10.1109/MMSP59012.2023.10337724

[14] Wang, L., Yeoh, B., Ng, J.W. (2022). Synthetic voice detection and audio splicing detection using se-res2net-conformer architecture. In 2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP), Singapore, Singapore, pp. 115-119. https://doi.org/10.1109/ISCSLP57327.2022.10037999

[15] Valente, L.P., de Souza, M.M., Da Rocha, A.M. (2024). Speech audio deepfake detection via convolutional neural networks. In 2024 IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS), Madrid, Spain, pp. 1-6. https://doi.org/10.1109/EAIS58494.2024.10569111

[16] Wu, Z., Kinnunen, T., Evans, N., Yamagishi, J., Alegre, F. (2015). Spoofing and countermeasures for speaker verification: A survey. Speech Communication, 66: 130-153. https://doi.org/10.1016/j.specom.2014.10.005

[17] Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., Plumbley, M.D. (2020). PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28: 2880-2894. https://doi.org/10.1109/TASLP.2020.3030497

[18] Wani, T.M., Amerini, I. (2023). Deepfakes audio

detection leveraging audio spectrogram and convolutional neural networks. In International Conference on Image Analysis and Processing (ICIAP), pp. 156-167. https://doi.org/10.1007/978-3-031-43153-1_14

[19] Liu, X., Kinnunen, T., Evans, N., Yamagishi, J. (2023). ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31: 2345-2359. https://doi.org/10.1109/TASLP.2023.3285283

[20] Tahaoglu, G., Baracchi, D., Shullani, D., Iuliani, M., Piva, A. (2025). Deepfake audio detection with spectral features and ResNext-based architecture. Knowledge-Based Systems, 323: 113726. https://doi.org/10.1016/j.knosys.2025.113726

[21] Rabhi, M., Bakiras, S., Di Pietro, R. (2024). Audio-deepfake detection: Adversarial attacks and defense countermeasure. Expert Systems with Applications, 250: 123941. https://doi.org/10.1016/j.eswa.2024.123941

[22] Sharafudeen, M., S S, V.C., J., A., Sei, Y. (2024). A blended framework for audio spoof detection with sequential models and bags of auditory bites. Scientific Reports, 14(1): 20192. https://doi.org/10.1038/s41598-024-71026-w

[23] Yi, J., Wang, C., Tao, J., Zhang, X., Zhang, C.Y., Zhao, Y. (2023). Audio deepfake detection: A survey. arXiv preprint arXiv:2308.14970. https://doi.org/10.48550/arXiv.2308.14970

[24] Li, X.F., Li, K., Zheng, Y.F., Yan, C., Ji, X.Y., Xu, W.Y. (2024). Safeear: Content privacy-preserving audio deepfake detection. In CCS '24: Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security, Salt Lake City, UT, USA, pp. 3585-3599. https://doi.org/10.1145/3658644.3670285

[25] Wang, R., Ye, D.P., Tang, L., Zhang, Y.M., Deng, J.C. (2024). AVT2-DWF: Improving deepfake detection with audio-visual fusion and dynamic weight strategies. IEEE Signal Processing Letters, 31: 1960-1964. https://doi.org/10.1109/LSP.2024.3433596

[26] Katamneni, V.S., Rattani, A. (2023). MIS-AVoiDD: Modality invariant and specific representation for audio-visual deepfake detection. In 2023 International Conference on Machine Learning and Applications (ICMLA), Jacksonville, FL, USA, pp. 1371-1378. https://doi.org/10.1109/ICMLA58977.2023.00207

[27] Yang, T., Sun, C., Lyu, S., Rose, P. (2025). Forensic deepfake audio detection using segmental acoustic features. Forensic Science International, 379: 112768. https://doi.org/10.1016/j.forsciint.2025.112768

[28] Di Pierno, A., Guarnera, L., Allegra, D., Battiato, S. (2025). End-to-end audio deepfake detection from raw waveforms: A RawNet-based approach with cross-dataset evaluation. arXiv preprint arXiv:2504.20923. https://doi.org/10.48550/arXiv.2504.20923

[29] El Kheir, Y., Das, A., Erdogan, E.E., Ritter-Guttierez, F., Polzehl, T., Möller, S. (2025). Two views, one truth: Spectral and self-supervised features fusion for robust speech deepfake detection. arXiv preprint arXiv:2507.20417. https://doi.org/10.48550/arXiv.2507.20417

[30] Uddin, K., Farooq, M.U., Khan, A., Malik, K.M. (2025). Adversarial attacks on audio deepfake detection: A benchmark and comparative study. arXiv preprint arXiv:2509.07132. https://doi.org/10.48550/arXiv.2509.07132