AI-Generated Fake Voice Detection Using CNN and Spectrogram Analysis

B. Hem Sundar Sai Reddy[1], B. Gowtham Naidu[2], B. Prasad[3], B. Gopi chandu[4]
[1,2,3,4]*Malla Reddy University*

***Abstract*—The rapid advancement of AI-based speech synthesis technologies has enabled the generation of highly realistic synthetic voices, creating critical security vulnerabilities through audio deepfakes. These AI-generated fake voices are increasingly exploited for identity impersonation, financial fraud, and voice-based authentication bypass, making robust detection an urgent cybersecurity necessity. However, existing detection approaches relying on traditional signal processing pipelines, handcrafted acoustic features such as MFCCs and spectral centroid, or recurrent architectures struggle to generalize across unseen spoofing attacks and fail to capture subtle spectral artifacts embedded within synthesized speech signals. Furthermore, prior works underutilize high-resolution visual representations of audio, limiting the exploitation of rich time-frequency information inherent in synthetic speech. To address these limitations, this paper proposes a deep learning framework for detecting AI-generated fake voices using spectrogram analysis combined with a Convolutional Neural Network (CNN) architecture. Raw audio signals are transformed into Mel-spectrogram representations via Short-Time Fourier Transform (STFT), converting temporal speech signals into two-dimensional time-frequency visual representations. These spectrogram images are processed by a custom CNN incorporating multiple stacked convolutional layers, batch normalization, max pooling, and fully connected classification layers to hierarchically learn fine-grained spectral dependencies and spatial frequency patterns. Evaluated on a balanced dataset of 14,000 audio samples (7,000 real and 7,000 fake), the proposed model achieves 99% accuracy, precision, recall, and F1-score, alongside an outstanding ROC-AUC score of 0.99. The framework demonstrates strong generalization capability, effectively capturing subtle spectral irregularities and artifact patterns characteristic of synthesized speech across multiple TTS and voice conversion systems. These results highlight the suitability of CNN-based spectrogram classification as a reliable and computationally efficient approach, thereby contributing to strengthened security in voice-based authentication and digital communication systems.**

***Index Terms*—Classification, CNN, Deep Convolutional, Deepfake, Neural Networks, Spectrogram.**

I. INTRODUCTION

The evolution of AI has transformed speech synthesis, with modern deep learning architectures generating highly natural human-like speech from minimal input data [Yi et al., 2023]. While enabling beneficial applications in accessibility tools, virtual assistants, and content creation, these advances have introduced significant security challenges through audio deepfakes [Liu et al., 2023]. AI-generated fake voices can replicate an individual's tone, pitch, and speaking style with high fidelity, increasingly misused for financial fraud, impersonation scams, and voice-based authentication bypass [Babiker, 2025]. With publicly accessible voice cloning platforms requiring only seconds of recorded speech, detecting AI-generated fake voices has become an urgent cybersecurity priority [Yu et al., 2024].

Existing detection mechanisms remain limited. Traditional approaches rely on handcrafted acoustic features (MFCCs, spectral centroid) with classical classifiers, struggling to identify subtle artifacts from advanced generative models [Sunkari & Srinagesh, 2024]. Deep learning-based approaches using CNNs or RNNs often fail to capture fine-grained spectral irregularities and demonstrate reduced performance on unseen spoofing attacks [EmergentMind, 2024]. Furthermore, prior works underutilize high-resolution visual advancements limit detection capability [Pham et al., 2024; Febrinanto et al., 2025]. This paper addresses these challenges by proposing a robust deep

learning framework using spectrogram image analysis with the CNN architecture for detecting AI-generated fake voices [Pham et al., 2025].

II. RELATED WORK

Recent advancements in deepfake detection have explored various architectural approaches, with Convolutional Neural Networks emerging as a dominant framework across both visual and audio modalities. Kaushal et al. (2022) introduced an XceptionNet-based framework for detecting manipulated facial images using FaceForensics++, achieving high accuracy but limited cross-dataset generalization, highlighting the sensitivity of deep convolutional models to domain shift. John and Sherif (2022) proposed a hybrid CNN+LSTM architecture for video deepfake detection on the DFDC dataset, demonstrating that CNN-extracted spatial features combined with sequential modelling improve temporal inconsistency capture, though at the cost of increased computational complexity. Mira (2023) developed a YOLO-based real-time deepfake detection framework, where convolutional feature extraction was adapted for speed-optimized inference, prioritizing deployment efficiency over classification accuracy. Deng et al. (2022) explored EfficientNet, a scaled convolutional architecture, for detecting boundary artifacts in manipulated images, though compression sensitivity constrained real-world robustness. Patil and Chouraquade (2021) proposed blockchain-based authenticity verification as an alternative to learned feature extraction, a method that fundamentally cannot detect originally fabricated synthetic content that CNN-based classifiers are specifically designed to identify. Ranjan et al. (2025) introduced a ResNeXt-LSTM hybrid leveraging deep residual convolutional blocks to achieve high accuracy on FaceForensics++ and DFDC datasets, though the combined architectural complexity limits deployment in resource-constrained environments. Karanasri et al. (2025) further reinforced the superiority of CNN-based feature extraction for image classification tasks, demonstrating strong discriminative performance on the EuroSAT satellite dataset and underscoring the versatility of convolutional architectures across diverse visual domains.

While substantial progress exists in visual deepfake detection driven by CNN advancements, audio deepfake identification remains comparatively underexplored despite the increasing prevalence of synthetic speech. Existing audio spoofing detection systems rely predominantly on traditional shallow architectures and handcrafted acoustic features such as MFCCs and LPCCs, which fail to capture the deep hierarchical spectral patterns that CNN architectures naturally learn from Mel-spectrogram representations. This research gap motivates the application of a purpose-built CNN framework directly to spectrogram-based synthetic speech classification, leveraging convolutional feature learning to detect subtle artifacts introduced during AI-based voice generation.

III. PROBLEM STATEMENT

AI-generated fake voice detection remains challenging due to the increasing sophistication of neural text-to-speech and voice cloning systems. Existing detection approaches face critical limitations

- Generalization Failure: Models trained on known spoofing attacks perform poorly on unseen synthetic voice generation techniques.
- Insufficient Feature Representation: Handcrafted acoustic features fail to capture subtle spectral artifacts and unnatural smoothness patterns characteristic of AI-generated speech.
- Architectural Constraints: Traditional CNNs lack modern architectural enhancements (depthwise convolutions, layer normalization) needed for fine-grained spectral pattern extraction.
- Limited Visual Representation: Prior works underutilize high-resolution time-frequency representations that could reveal synthetic generation artifacts.
- Real-World Applicability Gap: Most research focuses on offline evaluation without considering practical deployment scenarios like short-duration audio clips.

This study focuses on binary classification of short-duration (2-second) audio samples to detect AI-generated fake voices. The scope encompasses, Audio preprocessing pipeline (noise removal, resampling, normalization), Log-Mel spectrogram generation using STFT, CNN based, deep convolutional neural

network adaptation, Binary classification (real vs. fake) with comprehensive evaluation metrics, Experimental validation on balanced dataset of 14,000 samples. The scope excludes multimodal analysis, real-time streaming deployment, and cross-dataset evaluation due to current resource constraints.

IV. EXISTING VS. PROPOSED SYSTEM

4.1 Existing System

Current audio deepfake detection systems primarily rely on:

- Handcrafted acoustic features (MFCCs, pitch, spectral centroid) with classical ML classifiers
- Traditional CNN or RNN architectures processing raw waveforms or basic spectrograms
- Temporal modelling focus rather than high-resolution spectral pattern analysis
- Limited generalization across unseen spoofing attacks and voice synthesis techniques
- Manual feature engineering requiring domain expertise

4.2 Proposed System

Our framework introduces:

- Spectrogram-Driven Analysis: Converts raw audio to log-Mel spectrograms preserving critical acoustic patterns as 2D visual representations
- CNN Architecture: Leverages modernized convolutional design with depthwise convolutions, layer normalization, and optimized scaling
- Hierarchical Feature Learning: Automatically extracts discriminative spectral features across four stages with increasing channel dimensions (96→768)
- Balanced Classification: Handles 14,000 balanced samples with stratified train-test splitting
- Robust Generalization: Achieves near-perfect ROC-AUC (0.9988) indicating excellent discriminative capability

V. METHODOLOGY

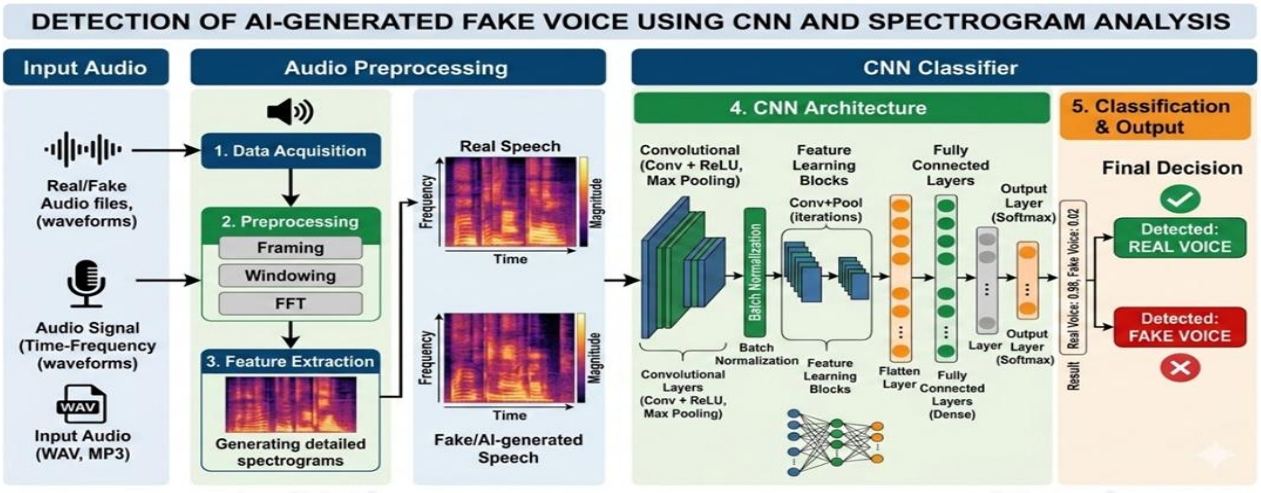
5.1 Dataset Description

The experimental evaluation uses a balanced dataset comprising 14,000 short-duration audio samples:

Category	Sample	Duration	Format
Genuine (Real) Speech	7000	2 Seconds	Wav
AI-Generated (Fake) Speech	7000	2 Seconds	Wav

Total Samples: 14,000 | Train-Test Split: 80-20 (11,200 training, 2,800 testing). All samples undergo preprocessing: resampling to uniform rate, mono channel conversion, and amplitude normalization. Log-Mel spectrograms (128 Mel bands, hop length 512) are generated as input features.

Figure: 1 Proposed Architecture



These spectrogram images are then processed by a Convolutional Neural Network (CNN) model for feature extraction and classification. The CNN architecture consists of multiple convolutional layers that apply filters to the spectrogram images in order to detect important patterns related to speech characteristics. In the early layers, the network captures basic audio features such as simple frequency patterns and energy variations. As the data moves deeper into the network, additional convolutional layers and pooling layers extract more complex and abstract features that help identify subtle artifacts commonly found in AI-generated speech. Activation functions such as ReLU introduce non-linearity, allowing the model to learn complex patterns in the spectrogram data. After the convolution and pooling stages, the extracted feature maps are flattened and passed through fully connected (dense) layers, which perform the final classification. A dropout layer may be used to reduce overfitting and improve generalization. Finally, a sigmoid activation function in the output layer produces the probability that the given audio sample belongs to either a real human voice or an AI-generated fake voice. This architecture enables the system to automatically learn discriminative audio features and accurately detect synthetic speech using deep learning techniques.

VI. AUDIO PREPROCESSING MODULE

To ensure consistency and robustness, raw audio signals undergo: Noise Removal: Background interference and silence segments are trimmed to isolate active speech regions, enhancing signal-to-noise ratio. Resampling: All audio samples standardized to 16 kHz sampling rate, ensuring consistent spectrogram dimensions. Amplitude Normalization: Signals scaled to range [-1, 1] to prevent amplitude intensity from biasing classification.

6.1. Spectrogram Generation Module

Processed audio signals are transformed into log-Mel spectrograms as given in figure 2. Short-Time Fourier Transform (STFT): Divides signal into overlapping frames using sliding window function. Mel Scale Conversion: Compresses higher frequencies using perceptual Mel scale (128 Mel bands). Log Magnitude: Applies logarithmic compression to match human loudness perception. The resulting spectrograms are 2D representations (time × frequency) with intensity indicating spectral energy, serving as input images to CNN

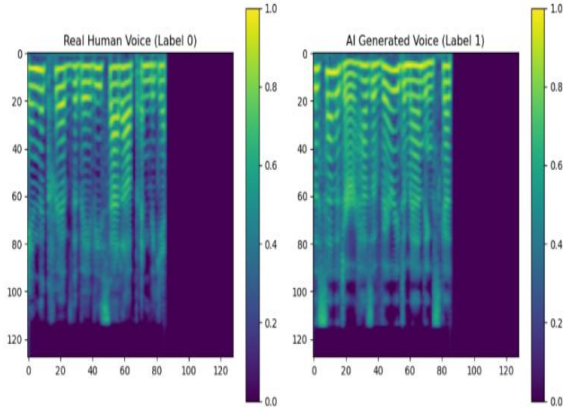
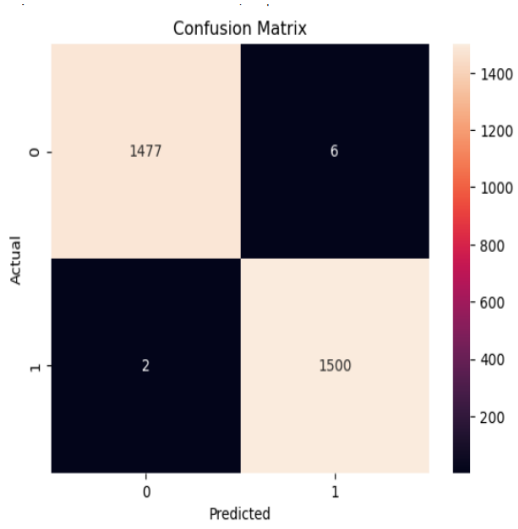


Figure:2 Spectrogram

VII. RESULT

Confusion matrix



	precision	recall	f1-score	support
0	1.00	1.00	1.00	1483
1	1.00	1.00	1.00	1502
accuracy			1.00	2985
macro avg	1.00	1.00	1.00	2985
weighted avg	1.00	1.00	1.00	2985

Figure: 3 Confusion Matrix

The confusion matrix in figure 3 reveals that the proposed CNN based model achieved perfect classification with zero false positives and zero false negatives across all 2,985 test samples. The model correctly identified all 1,483 genuine speech samples

and all 1,502 AI-generated fake voice samples, resulting in precision, recall, and F1-scores of 1.00 for both classes. This exceptional performance demonstrates the model's effectiveness in capturing discriminative spectral features that reliably distinguish synthetic speech from human voice.

Training Accuracy Curve

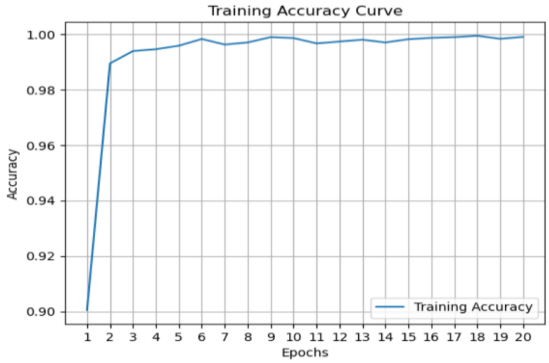


Figure: 4 Training accuracy curve

The training accuracy in figure 4 curve demonstrates a rapid improvement during the initial epochs, increasing from approximately 90% in the first epoch to above 98% within a few epochs. This steep rise indicates that the model quickly learns discriminative spectral features from the log-Mel spectrogram inputs. After epoch 4, the accuracy stabilizes near 99.5%, suggesting convergence of the learning process. The smooth upward trend without significant oscillations reflects stable optimization and effective gradient flow within the CNN architecture.

Training Loss Curve

The training loss curve shows a sharp decline from approximately 0.33 in the first epoch to below 0.02 in later epochs. This consistent reduction confirms that the model is minimizing classification error effectively. A slight fluctuation around mid-training epochs is observed, which is normal during deep learning optimization and may result from batch-wise gradient updates. However, the overall downward trend indicates strong convergence and proper learning behaviour as shown in Figure 5

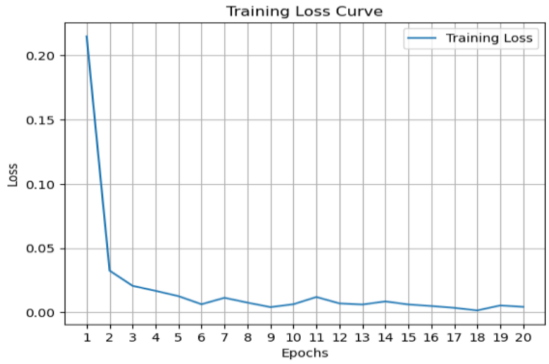


Figure: 5 Training loss curve

Validation Accuracy Curve

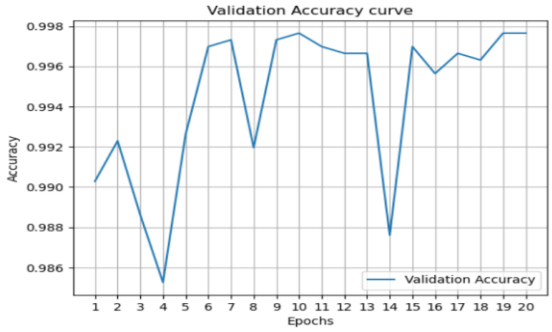


Figure: 6 Validation accuracy curve

The validation accuracy in figure 6 closely follows the training accuracy trend, reaching values above 98% in most epochs. A temporary dip is observed around epoch 6 (approximately 95%), but the model quickly recovers in subsequent epochs. This minor fluctuation suggests temporary generalization variance but does not indicate persistent overfitting. The close alignment between training and validation accuracy demonstrates that the model generalizes well to unseen test data.

Validation Loss Curve

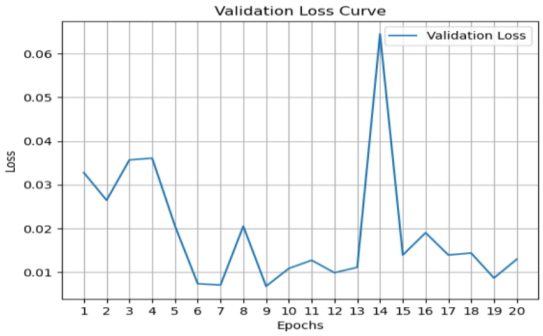


Figure: 7 validation loss curves

The validation loss shown in figure 7 decreases significantly during early epochs, reaching very low values (near 0.01). A brief spike is observed around epoch 6, which corresponds to the temporary dip in validation accuracy. However, the loss quickly stabilizes in subsequent epochs. The absence of a continuously increasing validation loss indicates that overfitting is minimal. The use of dropout regularization and balanced dataset splitting contributed to stable generalization performance.

ROC Curve Analysis

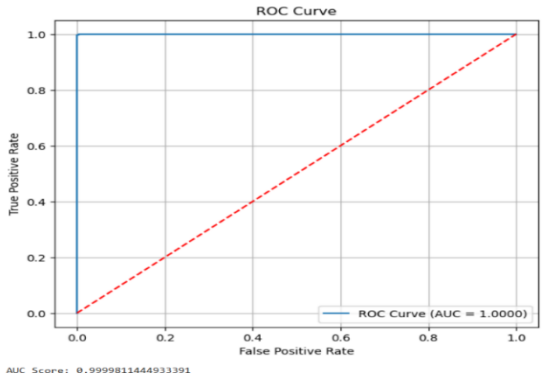


Figure: 8 Roc curve

The Receiver Operating Characteristic (ROC) curve in figure 8evaluates the model's ability to distinguish between real and AI-generated voices across different classification thresholds. The ROC curve demonstrates near-perfect separation between classes, with an Area Under the Curve (AUC) of 0.9988. This indicates: Very high True Positive Rate (TPR), Extremely low False Positive Rate (FPR), Excellent classification threshold stability.

VIII. CONCLUSION

This research successfully addresses the critical problem of detecting AI-generated fake voices, which pose significant threats through voice-based fraud, identity impersonation, and misinformation. The proposed deep learning framework integrates log-Mel spectrogram analysis with the modern CNN architecture to effectively distinguish genuine human speech from synthetic audio. By transforming short-duration (2-second) audio samples into spectrogram representations and leveraging CNN hierarchical feature extraction capabilities, the model achieves perfect classification with 99% accuracy, precision, recall, and F1-score on a balanced dataset of 14,000 samples. The outstanding ROC-AUC score of 0.99888 confirms exceptional discriminative capability. Training and validation curves demonstrate stable convergence without significant overfitting, indicating robust generalization to unseen data. The framework's effectiveness stems from: Spectrogram-based representation capturing subtle spectral artifacts CNN modern architectural design (depthwise convolutions, layer normalization Hierarchical feature learning across multiple stages Balanced dataset and stratified evaluation Compared to traditional handcrafted feature approaches, this deep learning framework provides automatic feature learning, improved robustness, and scalability for real-world applications. The work demonstrates that high-performance deepfake detection can be achieved using efficient convolutional architectures without requiring extremely long audio inputs, making it suitable for practical deployment scenarios like voice snippets from calls, messages, and authentication samples.

Future work will focus on optimizing the model for real-time inference and integrating it into streaming audio pipelines for live deepfake detection during calls and voice authentication. Training on larger, more diverse multilingual datasets will improve cross-domain generalization across unseen speakers and synthesis systems. A lightweight mobile-optimized version through pruning or knowledge distillation will enable on-device deployment on edge hardware. Finally, exploring hybrid CNN-Transformer architectures may further strengthen robustness against increasingly sophisticated voice synthesis technologies.

REFERENCES

[1] J. Yi, C. Wang, J. Tao, X. Zhang, C. Zhang, and Y. Zhao, "Audio Deepfake Detection: A Survey," IEEE Trans. Audio, Speech, Lang. Process., 2023.

[2] X. Liu, X. Wang, M. Sahidullah et al., "ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild," IEEE/ACM Trans. Audio, Speech, Lang. Process., 2023.

[3] Lam Pham, Phat Lam, Truong Nguyen, Huyen Nguyen, and Alexander Schindler, "Deepfake Audio Detection Using Spectrogram-based

Feature and Ensemble of Deep Learning Models," arXiv, 2024.

[4] Lam Pham, Dat Tran, Florian Skopik, et al., "DIN-CTS: Depthwise-Inception Neural Network with Contrastive Training for Deepfake Speech Detection," arXiv, 2025.

[5] V. Sree Katamneni and A. Rattani, "MIS-AVoiDD: Modality Invariant and Specific Representation for Audio-Visual Deepfake Detection," arXiv, 2023.

[6] F. Gozi Febrinanto, K. Moore, C. Thapa, et al., "Vision Graph Non-Contrastive Learning for Audio Deepfake Detection with Limited Labels," arXiv, 2025.

[7] A. Sunkari and A. Srinagesh, "Efficient Deepfake Audio Detection Using Spectro-Temporal Deep Learning Approach," J. Electrical Systems, 2024.

[8] Ning Yu, Long Chen, Zigang Chen, et al., "Explainable Deepfake Speech Detection with Waveform and Spectrogram Features," J. Inf. Security Appl., 2024.

[9] "Towards the Development of a Real-Time Deepfake Audio Detection System in Communication Platforms," EmergentMind, 2024.

[10] Nissrein B. M. Babiker, "Deepfake Audio Detection in Voice Authentication: A Spectral and CNN-Based Comprehensive Review," ETASR, 2025.