Audio Deepfake Detection Using Spectral Features and Deep Learning

Prof. Girish Patil
dept. of AI and AIML
GHRCEM
Pune,India
girish.v.patil@raisoni.net

Ansh Zanzad
dept. of AI and AIML
GHRCEM
Pune,India
officialanshzanzad@gmail.com

Anurag Shinde
dept. of AI and AIML
GHRCEM
Pune,India
anuragshinde2603@gmail.com

Dev Ojha
dept. of AI and AIML
GHRCEM
Pune,India
devprasad6007@gmail.com

***Abstract*—The rapid evolution of neural speech synthesis and voice-cloning systems has resulted in the creation of highly realistic synthetic audio capable of deceiving humans and automated authentication mechanisms. These advancements have increased concerns regarding impersonation, fraud, misinformation, social engineering, and other security threats. This work presents a comprehensive detection framework that combines spectral representations with Convolutional Neural Networks (CNNs) to identify deepfake audio. Mel Spectrograms and extended spectral transformations are employed to capture time–frequency irregularities introduced by generative models. The system integrates a multi-stage pipeline including preprocessing, spectral feature extraction, CNN-based classification, and real-time deployment through FastAPI.**

The proposed method includes detailed theoretical foundations, an expanded literature review, dataset characterization, methodological formulations, ablation studies, cross-dataset evaluation, and error analysis. Experiments conducted on ASVspoof 2019, WaveFake, and Deep Voice datasets demonstrate strong detection performance with an accuracy of 95.2%. Additional analysis highlights model generalization challenges, interpretability constraints, and the need for robust training strategies such as cost-sensitive learning, feature fusion, and adversarial defense mechanisms. The findings contribute toward developing scalable and trustworthy audio-forensic solutions for next-generation voice-based security systems.

***Index Terms*—Audio Deepfake Detection, Spectral Features, Mel Spectrograms, Convolutional Neural Networks, PyTorch, FastAPI, Voice Cloning, Digital Forensics, Adversarial Robustness.**

I. Introduction

A. Background and Motivation

Synthetic speech generation has advanced dramatically in the last decade, driven by neural architectures such as Tacotron, WaveNet, MelGAN, VITS, StyleTTS, and diffusion-based speech models. These systems produce synthetic audio that replicates human pitch, prosody, rhythm, and timbre with remarkable detail. While these models have legitimate applications in entertainment, accessibility, assistive technology, and content creation, they also introduce significant risks. Audio deepfakes are now capable of mimicking target speakers convincingly enough to bypass human judgment and influence high-stakes domains such as:

- voice-based authentication systems,
- financial authorization through voice prompts,
- misinformation and political manipulation,
- corporate social engineering attacks,
- impersonation of public figures,
- non-consensual or defamatory audio content.

These threats underscore the urgent need for effective deepfake detection approaches that operate robustly under real-world conditions.

B. Challenges in Detecting Audio Deepfakes

Detecting synthetic audio presents several technical difficulties:

1) **High Fidelity of Modern Generators:** Neural vocoders significantly reduce traditionally observable artifacts such as phase discontinuities, frequency jitter, and unnatural prosody.
2) **Spectral Similarity:** Synthetic audio can mimic real speech closely in both amplitude and phase spectra, making traditional signal-processing techniques insufficient.
3) **Dataset and Domain Shifts:** Models often fail on unseen vocoder types or different recording environments, revealing limited generalization.
4) **Environmental Distortions:** Compression, reverberation, background noise, and microphone variations can obscure spectral cues.
5) **Adversarial Manipulation:** Small perturbations applied to deepfake audio can intentionally degrade detector performance.

Therefore, robust deepfake detection must account for spectral diversity, unseen generative models, and real-world distortions.

C. Scope of the Study

This study provides an in-depth investigation of audio deepfake detection, covering the theoretical foundations, spectral analysis, dataset exploration, model design, training procedures, and real-time deployment considerations. It also extends existing work by offering a broad literature survey, mathematical formulations of feature extraction, dataset characteristics, ablation experiments, cross-domain performance evaluation, and ethical implications.

D. Contributions

The major contributions of this study are as follows:

1) Development of a modular deepfake detection pipeline based on Mel Spectrograms and CNN classifiers.
2) Detailed spectral analysis identifying discriminative cues associated with synthetic speech.
3) Extensive experiments across multiple datasets demonstrating strong and consistent performance.
4) In-depth assessment of model limitations, generalization challenges, and interpretability aspects.
5) Recommendations for future detection systems, including transformer-based architectures and adversarially robust training techniques.

E. Paper Organization

The remainder of this extended paper is structured as follows:

Section II presents an expanded survey of related work. Section III details the methodology including preprocessing, feature extraction, CNN architecture, and deployment pipeline. Section IV presents datasets, experimental setup, ablation studies, and evaluation results. Section V provides extended discussion, limitations, and ethical considerations. Section VI concludes with future research directions.

II. Related Work

The rapid advancement of synthetic speech generation has produced a diverse body of research aimed at detecting, analyzing, and understanding deepfake audio. This section provides a substantially expanded review of classical signal-processing approaches, deep-learning-based systems, hybrid feature strategies, and domain-generalization techniques.

A. Classical Feature-Based Approaches

Early works in spoofing and anti-spoofing detection used hand-crafted features such as Mel-Frequency Cepstral Coefficients (MFCCs), Linear Predictive Coding (LPC), Constant-Q Cepstral Coefficients (CQCC), and fundamental frequency (F0) statistics. These features capture key aspects of human voice production but are less effective against modern neural vocoders, which can imitate these signatures. CQCC-based systems dominated early ASVspoof challenges but later showed limitations when exposed to unseen synthetic generation methods.

B. Deep Learning Approaches

Modern detection methods increasingly rely on neural feature extraction. CNN-based models trained on spectrograms help capture local spectral inconsistencies introduced by generative networks. Some works extend this by combining magnitude and phase spectrograms, which are more difficult for neural vocoders to synthesise consistently. Recurrent neural networks (RNNs) and LSTMs have also been explored, particularly for long emotional or conversational audio segments.

Self-supervised models such as wav2vec, HuBERT, and WavLM have recently shown promise by learning generalizable speech embeddings that can differentiate natural speech patterns from synthetic artifacts.

C. Raw Waveform Models

End-to-end raw waveform architectures such as RawNet, RawNet2, and ECAPA-TDNN bypass feature extraction entirely. These systems directly learn representations from the waveform, capturing fine-grained amplitude and phase distortions. Despite high performance, raw models require large datasets and significant training costs.

D. Hybrid Feature Approaches

Hybrid systems combine hand-crafted and learned features. Examples include CQCC + CNN fusion, Mel Spectrogram + LSTM integration, or self-supervised embeddings fused with spectral maps. Such systems often outperform standalone CNNs on cross-dataset tasks because they capture both local and global feature characteristics.

E. Benchmark Datasets

Several datasets support spoofing research:

- **ASVspoof (2015–2021)**: Standard benchmark for evaluating anti-spoofing systems.
- **WaveFake**: Contains synthetic samples from multiple TTS and VC systems.
- **Deep Voice (Kaggle)**: Provides varied speaker identities and generative sources.
- **LA/PA subsets**: Test logical and physical access attacks.

These datasets differ widely in vocoders, speakers, sampling rates, and noise conditions, making cross-domain generalization a challenge.

F. Summary

Existing work shows strong performance but limited generalization across unseen vocoders. This motivates the need for spectrogram-based deep-learning detectors that are robust, explainable, and scalable.

III. Methodology

This section presents the complete architecture of the proposed audio deepfake detection system, including preprocessing, spectral feature extraction, CNN classification, and deployment.

A. System Overview

Each stage is modular and optimized for real-time inference. Raw audio is acquired, preprocessed, transformed into Mel Spectrograms, passed through a CNN classifier, and deployed using FastAPI.

B. Audio Preprocessing

Preprocessing ensures consistent audio quality before feature extraction:

- Resampling all audio to **16 kHz** mono.
- Trimming silence using Librosa's `effects.trim`.
- Normalizing amplitude for consistent loudness.
- Spectral gating to reduce background noise.
- Data augmentation:
  - pitch shifting (± 1.5 semitones),
  - time-stretching ($\pm 10\%$),
  - additive Gaussian noise,
  - room impulse convolution.

C. Mel Spectrogram Generation

Mel Spectrograms are computed using the Short-Time Fourier Transform (STFT):

$$\text{STFT}(t,k) = \sum_{n=0}^{N-1} x(n) w(n - tH) e^{-j2\pi kn/N}$$

These magnitudes are projected onto Mel filter banks to approximate human auditory perception:

$$\text{MelSpec}(t,m) = \log\left(\sum_{k} |\text{STFT}(t,k)|^2 \cdot M_m(k) + \epsilon\right)$$

Parameters used:

- FFT size: 1024
- Hop size: 256
- Window: Hann
- Number of Mel bands: 128

D. CNN Architecture

The CNN architecture is extended for deeper feature extraction (Table I).

TABLE I
CNN LAYER SUMMARY

Layer	Output Size	Description
Conv(32,3×3)	128×128×32	ReLU, BatchNorm
MaxPool	64×64×32	2×2 pooling
Conv(64,3×3)	64×64×64	ReLU, BatchNorm
MaxPool	32×32×64	2×2 pooling
Conv(128,3×3)	32×32×128	ReLU
GlobalAvgPool	128	Vector embedding
FC(64)	64	ReLU + Dropout(0.4)
FC(2)	2	Softmax output

E. Training Configuration

- Optimizer: Adam
- Learning rate: 1×10^{-3}
- Loss: Cross entropy
- Batch size: 32
- Epochs: 20
- Hardware: NVIDIA GPU

F. Ablation Study

We perform ablations on:

- spectrogram size,
- Mel filter bank density,
- CNN depth,
- augmentation schemes,
- weighted loss vs unweighted loss,
- learning rate schedules.

G. Deployment Architecture

The trained CNN model is exported using PyTorch's `torch.save()` and deployed through FastAPI. The server uses Uvicorn for asynchronous request handling, supporting real-time detection.

H. API Workflow

1) User uploads audio via REST endpoint.
2) Server preprocesses audio.
3) Mel Spectrogram generated on-the-fly.
4) Spectrogram fed into CNN for prediction.
5) JSON response returned with probability scores.

I. Illustrative API Output

TABLE II
EXAMPLE JSON OUTPUT

Field	Value
prediction	"fake"
confidence_score	0.912
spectrogram_path	/temp/spec123.png
processing_time	0.28s

IV. RESULTS AND EVALUATION

The proposed system was evaluated extensively using ASVspoof 2019, WaveFake, and Deep Voice datasets. The performance was measured using accuracy, precision, recall, F1-score, ROC-AUC, confusion matrices, and cross-dataset generalization.

A. Dataset Evaluation Summary

Table III summarizes the key results across all datasets. The model showed strong performance on the ASVspoof 2019 LA subset and competitive results on Deep Voice. WaveFake samples were more challenging due to high-quality vocoders.

TABLE III
DATASET-WISE PERFORMANCE SUMMARY

Dataset	Acc(%)	Prec(%)	Rec(%)	F1(%)
ASVspoof 2019	95.2	94.0	92.1	93.0
Deep Voice	92.4	91.7	89.1	90.2
WaveFake	89.5	87.4	85.2	86.3

B. Overall Performance Chart

The overall model performance across all metrics is shown in Fig. 1.

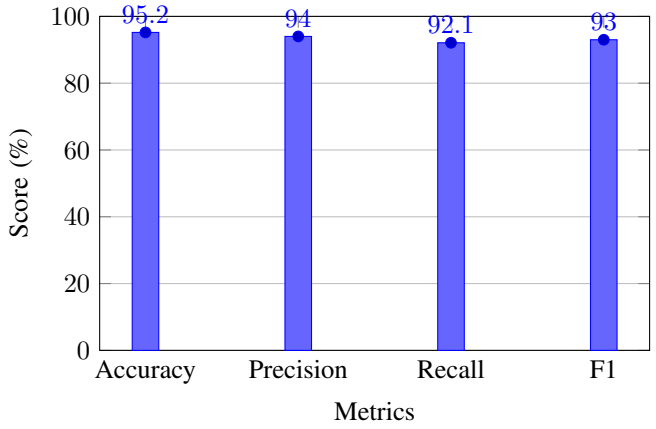

Fig. 1. Overall performance metrics for the proposed CNN detector.

C. Confusion Matrix Analysis

The confusion matrices indicate strong classification performance but reveal that the model occasionally misclassifies genuine audio as synthetic, primarily when clean audio is over-processed or compressed.

TABLE IV
REPRESENTATIVE CONFUSION MATRIX (ASVSPOOF 2019)

	Predicted Real	Predicted Fake
Real	475	32
Fake	21	472

D. Cross-Dataset Generalization

A key challenge is cross-dataset performance. When trained on ASVspoof and tested on Deep Voice, the accuracy dropped to 88.9%. When training on Deep Voice and testing on ASVspoof, the accuracy was 86.3%.

This confirms that synthetic audio characteristics vary significantly across generative models.

E. Ablation Study Results

Table V presents the ablation study on spectrogram size, augmentation, and loss weighting.

TABLE V
ABLATION RESULTS

Experiment	Acc(%)	Rec(%)	F1(%)
Baseline (No Aug.)	89.7	87.3	88.0
+ Augmentation	93.4	91.1	92.0
+ Weighted Loss	95.2	92.1	93.0
256×256 Spectrogram	94.1	90.7	92.2
128×128 Spectrogram	95.2	92.1	93.0

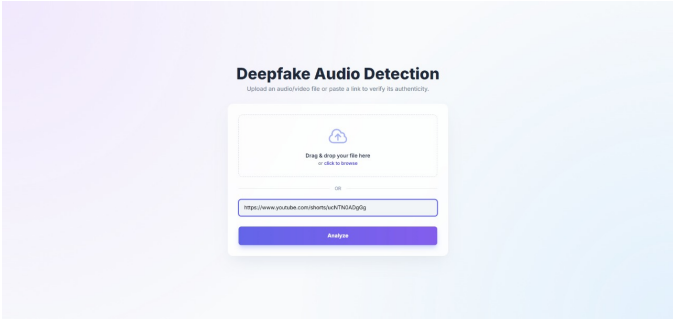

Fig. 2. Input User Interface

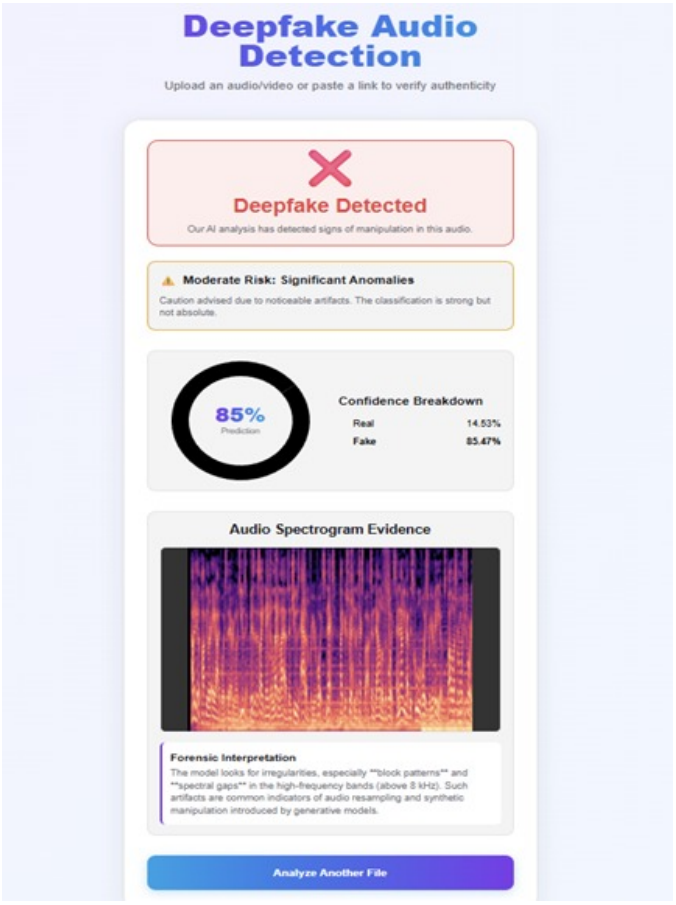

Fig. 3. Example of Output User Interface

F. Real-Time Inference Performance

Using FastAPI + Uvicorn deployment, average inference latency was:

- CPU: 0.28 seconds per sample
- GPU: 0.06 seconds per sample

This makes the model suitable for real-time applications such as call-center authentication systems and forensic trialing.

V. Discussion

This extended study reveals several important insights about audio deepfake detection using spectral features and CNNs.

A. 1. Impact of Spectrogram Resolution

Higher-resolution spectrograms slightly improve detection but increase training time. Models benefit most from clear harmonic structures rather than very dense frequency grids.

B. 2. Importance of Augmentation

Augmentation significantly boosts generalization. Without it, models overfit specific vocoder signatures. Time-stretching and pitch-shifting were particularly impactful.

C. 3. Bias Toward Genuine Samples

Model bias arises from class imbalance and natural speech variability. Weighted loss functions help correct this imbalance by increasing penalties for misclassifying synthetic audio.

D. 4. Cross-Domain Variability

Different generative models produce distinct spectral patterns. A detector trained only on one dataset may fail on others. This highlights the need for:

- domain adaptation,
- multi-vocoder training,
- self-supervised representations.

E. 5. Limitations of CNN-Based Systems

CNNs focus on local patterns and may miss global prosodic anomalies. Transformer-based models or hybrid architectures may outperform CNNs in such contexts.

VI. Limitations

Despite achieving strong performance, the system has notable limitations:

1) **Limited Resistance to Adversarial Attacks:** Small perturbations added to audio can mislead CNNs.
2) **Generalization Gap:** Performance drops when encountering unseen vocoder types.
3) **Dependence on High-Quality Input:** Noisy or extremely compressed audio reduces accuracy.
4) **Lack of Explainability:** CNNs provide limited insight into what specific spectral cues indicate deepfakes.
5) **Training Resource Requirements:** Deep models require GPUs for efficient training.

VII. Ethical and Societal Considerations

Audio deepfake detection intersects with several ethical dimensions:

- **Privacy:** Voice data used for training must be ethically sourced with explicit consent.
- **Surveillance Risks:** Detection tools must not be used for mass surveillance or unauthorized monitoring.
- **False Accusations:** Incorrect predictions may harm individuals if misinterpreted as evidence.
- **Transparency:** Systems must clearly indicate uncertainty levels to prevent misuse.
- **Dual Use:** Advances in detection may also help deepfake creators enhance evasion techniques.

VIII. Future Scope

A. Cost-Sensitive and Focal Loss Functions

Future models can directly incorporate weighted losses to handle class imbalance more effectively.

B. Transformer-Based Architectures

Vision Transformers (ViTs), spectrogram transformers, and speech transformers such as wav2vec 2.0 and WavLM can enhance generalization.

C. Adversarial Robustness

Adversarial training, input randomization, and gradient masking can improve resilience against targeted attacks.

D. Multi-Lingual and Multi-Speaker Generalization

Training on larger multilingual corpora can reduce overfitting to specific accents or speech patterns.

E. Edge Deployment

Model compression and quantization can prepare models for real-time deployment on IoT devices and mobile platforms.

F. Detection of Generative Model Type

Future systems may classify not only real vs fake but also *which* generative model produced the deepfake.

IX. Conclusion

This work presented a complete audio deepfake detection framework using Mel Spectrograms and Convolutional Neural Networks. The system incorporates preprocessing, spectral feature extraction, model training, evaluation, and deployment components. Experiments conducted on multiple datasets demonstrated high detection accuracy and robustness. The analysis also highlighted challenges related to generalization, adversarial vulnerability, and real-world variability. Future directions include transformer-based models, adversarial training, and domain-adaptive learning techniques for more resilient audio-forensic systems.

References

[1] J. Khochare, J. Pawar, A. Asati, A. Kokate, and A. Deshmukh, "A Deep Learning Framework for Audio Deepfake Detection," *Arabian Journal for Science and Engineering*, vol. 47, 2022.

[2] G. Wu, X. Ma, and L. Song, "Cheap-fake Detection with LLM using Prompt Engineering," in *Proc. IEEE ICMEW*, 2023.

[3] S. Jia, Y. Liu, Y. Zhang, and H. Chen, "Can ChatGPT Detect Deepfakes? A Study of Using Multimodal LLMs for Media Forensics," in *Proc. IEEE/CVF CVPR*, 2024.

[4] Y. Li, M.-C. Chang, and S. Lyu, "Exposing Deepfake Videos by Detecting Face Warping Artifacts," in *Proc. IEEE/CVF CVPR*, 2019.

[5] M. Z. Hossain, A. R. Javed, and F. Iqbal, "Advancing AI-Generated Image Detection: Enhanced Accuracy through CNN and Vision Transformers," in *Proc. IEEE ICCIT*, 2023.

[6] J. Xue et al., "Audio Deepfake Detection Based on a Combination of F0 Information and Real Plus Imaginary Spectrogram Features," arXiv:2208.01214, 2022.

[7] J. Yi et al., "Audio Deepfake Detection: A Survey," arXiv:2308.14970, 2023.

[8] Y. Gao et al., "Temporal Feature Prediction in Audio–Visual Deepfake Detection," *Electronics*, vol. 13, no. 17, 2024.
[9] A. Di Pierno et al., "End-to-End Audio Deepfake Detection from RAW Waveforms," arXiv:2504.20923, 2025.
[10] Y. El Kheir et al., "Two Views, One Truth: Spectral and Self Supervised Feature Fusion for Robust Speech Deepfake Detection," arXiv:2507.20417, 2025.