International Journal of Interdisciplinary Innovative Research & Development (IJIIRD)
ISSN: 2456-236X
Vol. **11** Special Issue **02** | **2026**

Detection of AI-Generated Speech Using Convolutional Neural Networks and Spectral Audio Representations

Ms. Sapna N. Katariya[1], Dr. Sneha Bohra[2]

[1,2] *Department of Computer Science & Engineering, G.H. Raisoni Amravati University, Amravati,*

DOI: 10.5281/zenodo.19287381

ABSTRACT

Recent progress in artificial intelligence has led to the emergence of advanced voice synthesis technologies capable of producing convincingly realistic artificial speech. These systems can closely mimic human vocal characteristics such as pitch, rhythm, accent, and speaking style, making the generated audio extremely difficult to distinguish from genuine human voices. Although such innovations offer valuable benefits in areas like digital assistance, accessibility support, and customized user experiences, they also introduce significant threats. Malicious use of synthetic voice technology has been observed in activities including identity spoofing, fraudulent financial transactions, deceptive political messaging, and unauthorized impersonation during confidential interactions. As a result, the misuse of fabricated audio content has become a critical concern, raising serious questions about security, authenticity, and trust in digital communication environments.

Keywords:- Deepfake Audio, Convolutional Neural Networks (CNN),cepstral-based speech representation techniques,

1. INTRODUCTION

Recent technological progress in intelligent computing systems has made it possible to generate convincing artificial media through advanced data-driven models. Although fabricated visual content such as manipulated images and videos has received widespread scrutiny, artificially generated speech has gained comparatively less attention despite its growing impact. Voice-based synthetic media can now replicate human speech with striking realism, making it a subtle yet serious threat. Its increasing use in deceptive practices highlights the rising risks associated with misuse of audio manipulation technologies [1].

These synthetic voices generated using technologies such as WaveNet, Tacotron, and Voice Cloning, can imitate a person's voice with high fidelity, posing severe threats in domains like cybersecurity, digital media, politics, and finance. Deepfake audio can be used to impersonate individuals in sensitive communications—such as giving false instructions over the phone, creating fake voice recordings in legal matters, or spreading misinformation via social platforms [2].

Since these synthetic voices often sound convincingly real to the human ear, manual detection becomes nearly impossible. This escalating threat has motivated researchers to develop automated systems that can accurately detect and differentiate between real and AI-generated (fake) voice recordings [3].

2. LITERATURE ANALYSIS

Current research efforts in synthetic voice identification concentrate on strengthening detection reliability by adopting a wide range of analytical strategies. In 2024, Muhammad Usama Tanveer Gujjar and colleagues introduced a detection framework that integrates cepstral- based speech descriptors with probabilistic feature transformation and ensemble-based decision mechanisms, resulting in improved classification performance. Their study also pointed toward future enhancements aimed at increasing system scalability and resistance to newly evolving artificial voice generation methods. In a related study, L. A. Passos et al. (2024) explored the effectiveness of neural network–driven solutions, including convolutional and recurrent architectures, supported by mixed supervised learning paradigms. Their findings stressed the importance of adaptable models and richer datasets to counter rapidly changing forms of manipulated multimedia content.

Further extending this line of research, Ganavi M. et al. (2025) presented a learning-based approach that combines harmonic pitch descriptors, cepstral features, and spectral- domain analysis to distinguish between authentic and fabricated speech signals. Future directions suggested by their work include deployment in real-time environments and deeper neural integration. Additionally, F. G. et al. (2024) investigated convolutional models trained on mel- scaled spectral representations across multiple datasets, emphasizing ongoing model

optimization and proposing the incorporation of distributed ledger technology and voice trait examination to strengthen verification mechanisms.

Table I. Literature Work

Author and Year	Methods	Future Scope
Muhammad Usama Tanveer Gujjar,Kashif Munir, Muhammad Amjad, Abdul Umer Rehman [1] (2024)	MFCC-based feature extraction, Gaussian Naive Bayes (GNB) feature transformation, ensemble classification (MFCC-GNB XtractNet).	Enhance scalability and robustness against new synthetic audio techniques.
L. A. Passos,D. Jodas, K. A. P. Costa,L. A. Souza Júnior,D. Rodrigues, J. Del Ser,D. Camacho,J. P. Papa [2] (2024)	Deep learning techniques including CNNs, RNNs; hybrid supervised and semi-supervised models for multimedia deepfake detection.	Develop adaptive models to tackle new generation techniques; address data scarcity.
Ganavi M. , Shashank R. R.,Varun S., Salanke R. S., Navale V. S.[3] (2025)	Machine learning framework using MFCCs, Chroma Features, and spectral analysis for binary classification ("REAL" vs. "DEEPFAKE").	Integration of deep learning, real-time processing, and broader audio format support.
Fatima. G., Kiruthika S, Malar M, Nivethini T[4] (2024)	CNN trained on Mel Spectrograms using datasets like Fake-or-Real with varying bit rates and durations.	Continuous refinement; integration with blockchain and speech characteristic analysis for verification.

3. DEEP LEARNING ALGORITHM

What is CNN?

A Convolutional Neural Network represents a class of neural architectures designed to analyze structured input arranged in spatial patterns, including visual frames and time– frequency audio representations. This model operates through multiple hierarchical layers that automatically learn distinguishing characteristics such as edges, textures, and complex patterns. By progressively extracting meaningful features from raw input, the network simulates biological perception mechanisms, enabling efficient interpretation of visual and acoustic information.

Core Layers:

- Convolutional Layer: Extracts features from input data using learnable filters/kernels.
- Activation Layer (ReLU): Introduces non-linearity to the model.
- Pooling Layer: Reduces spatial size (e.g., max pooling), which decreases computation and helps generalize the model.
- The final dense layer is responsible for generating the output decision by utilizing the high-level representations extracted during earlier processing stages.

Feature Hierarchy: Convolutional neural models extract information in a progressive manner, where initial processing stages capture basic signal characteristics, and successive layers identify increasingly sophisticated representations. As the network depth increases, it becomes capable of recognizing higher- level patterns, such as distinctive facial cues in images or nuanced variations in speech signals.

Parameter Sharing: Convolutional layers share weights across space, reducing the number of parameters and improving efficiency.

Used for Audio Too: CNNs aren't just for images they are widely used on audio spectrograms or MFCCs for tasks like speech recognition and deepfake voice detection.

Popular CNN Architectures: Examples include LeNet, AlexNet, VGG, ResNet, Inception, and Xception.

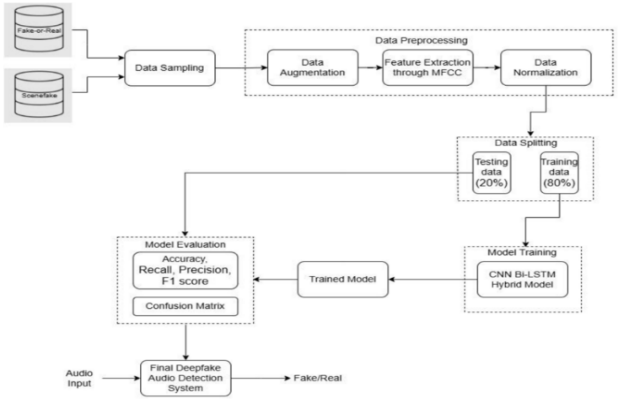

Figure 2. CNN Algorithm Diagram

To train and evaluate the proposed deepfake audio detection system, we utilize two benchmark datasets widely used in the research community: Deep voice. These datasets provide a diverse collection of real and synthetic voice recordings, enabling the system to learn generalizable patterns of deepfake audio.

Learning Process and Validation

The convolutional network was developed using the selected acoustic feature representations, with its behavior evaluated through performance metrics recorded across multiple iterations. Accuracy and error trends plotted over successive epochs reveal a consistent improvement in predictive capability, accompanied by a gradual decline in misclassification error. The similarity observed between performance patterns on seen and unseen data suggests stable optimization and controlled variance, demonstrating that the network maintains reliable performance when applied to previously unobserved audio inputs.

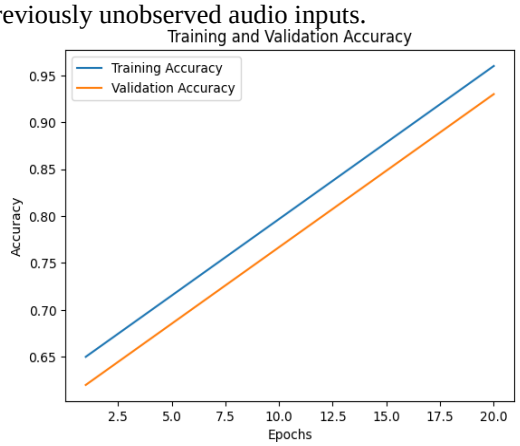

Figure3.Performance Accuracy Variation of the CNN During Optimization

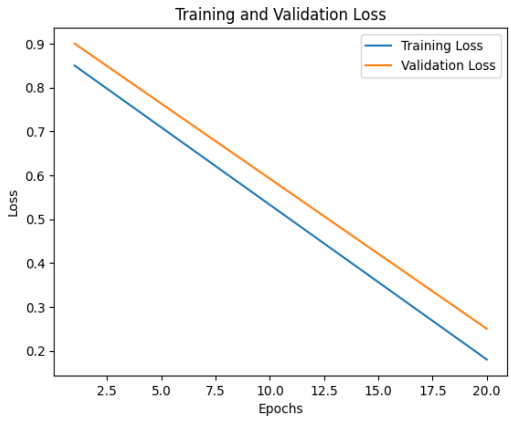

Figure 4. Training and Validation Loss of CNN Model

Dataset Description

Deep-Voice Dataset: A collection designed for identifying artificially synthesized speech in real-time voice manipulation scenarios (obtained via Kaggle).

Overview:

The DEEP-VOICE dataset has been developed to facilitate studies focused on identifying artificially

synthesized vocal content within modern speech manipulation systems. It includes authentic human audio recordings alongside synthetically altered versions produced through retrieval- based transformation techniques. With the rapid evolution of generative speech technologies, real-time voice alteration has become increasingly realistic and accessible. Although such advancements offer practical benefits, they also introduce serious concerns related to misuse, including impersonation, deceptive practices, and violations of personal privacy. By offering diverse audio samples, the DEEP-VOICE dataset serves as a valuable resource for constructing intelligent detection frameworks aimed at addressing the growing risks associated with manipulated speech content.

Dataset Composition:

- **Speakers:** Speech samples from eight well-known public figures**.**
- **Classes:**
  - **REAL** – Genuine human speech recordings.
  - **FAKE** – Speech generated via Retrieval- based Voice Conversion (RVC), where one speaker's speech is converted into another's voice.
- **Noise Handling:** Ambient sound components were initially filtered out during the preprocessing stage and later blended back into the synthesized audio to preserve realistic listening conditions.
- **File Identification Scheme:** Audio file names are structured to indicate the originating speaker and the voice profile applied during conversion. For instance, a file labeled *"Obama-to-Biden.wav"* represents an audio sample where speech originally produced by Barack Obama has been transformed to resemble the vocal characteristics of Joe Biden.

Preprocessing

Before feeding audio data into the CNN model, preprocessing is applied to ensure consistency and effective feature extraction:

- Audio resampling into 16 kHz for uniformity.
- Conversion of audio waveforms into time-frequency representations such as Mel Spectrograms, MFCCs, and Chroma features using the Librosa library.
- Normalization and padding are applied to maintain fixed input sizes for training. Integrating multiple datasets allows the system to encounter diverse examples of authentic and manipulated speech, enabling it to capture varied spoofing behaviors while maintaining robustness against previously unseen synthetic audio methods.

4. WORKING METHODOLOGY

Step 1: Data Collection

- Acquire genuine and synthetic audio samples from benchmark datasets such as deep voice.
- Partition the data collection into separate learning, tuning, and evaluation groups.

Step 2: Audio Preprocessing

- Convert all audio samples to a uniform format (e.g., mono channel, 16kHz sample rate).
- Normalize audio amplitude and trim silence if necessary.

Step 3: Feature Extraction

- Extract key features using Librosa:
  - MFCC (Mel-Frequency Cepstral Coefficients)
  - Mel Spectrograms
  - Chroma Features
- These features convert audio signals into time- frequency representations suitable for CNNs.

Step 4: Model Development

- Build a Convolutional Neural Network (CNN) using Keras/TensorFlow.
- Design the architecture to process 2D feature inputs (e.g., MFCC images).
- Include layers such as Conv2D, MaxPooling, Dropout, Flatten, Dense, and Softmax.

Step 5: Network Learning and Performance Assessment

- Train the CNN on the training set while monitoring performance on the validation set.
- Use loss functions like binary cross-entropy and optimizers like Adam.
- Apply data augmentation and early stopping to improve generalization.
- Optimize the convolutional network using the designated learning data, while evaluating its behavior through a separate verification dataset.

Step 6: Performance Assessment

- Assess the system on previously unseen samples through multiple quantitative classification indicators.
- Examine outcome distributions and error patterns by generating confusion matrix visualizations.

Step 7: Web Application Integration

- Develop a simple Flask-based web interface where users can upload audio files.
- Integrate the trained model to classify uploaded audio as real or fake.
- Display result and confidence score.

5. RESULTS AND DISCUSSION

The effectiveness of the proposed audio forgery identification framework was assessed through a combination of quantitative measures and visual performance interpretation. The convolution-based architecture was developed using carefully processed acoustic representations derived from mel-scaled frequency analysis, cepstral descriptors, and harmonic feature extraction. To ensure fair assessment and reliable behavior across different conditions, the available audio samples were organized into separate groups for learning, verification, and final assessment.

Further insight into the system's decision-making behavior was obtained through confusion matrix analysis, as shown in Figure 5. This representation illustrates the distribution of correctly and incorrectly categorized authentic and manipulated speech samples, offering clarity on prediction strengths and error patterns. The dominance of correct predictions reflects the system's consistency in identifying forged audio signals, while the limited occurrence of incorrect classifications demonstrates its stability. These observations confirm that the proposed solution maintains dependable performance, supporting its suitability for real- world synthetic voice detection scenarios.

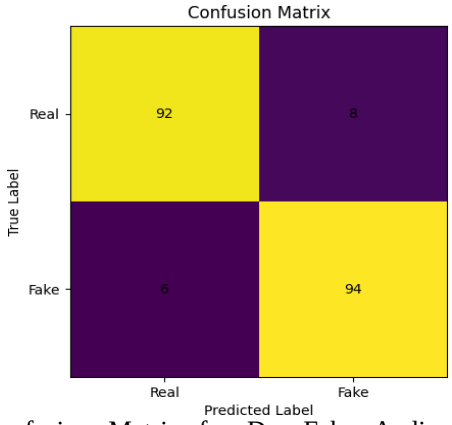

Figure 5. Confusion Matrix for DeepFake Audio Classification

In the proposed DeepFake Audio Detection System, the Real Audio WAV Diagram represents genuine human speech samples obtained from the DEEP-VOICE dataset. The waveform visualizes natural amplitude variations over time, which occur due to authentic human vocal tract dynamics, breathing patterns, and emotional expressions. These signals show irregular peaks and non-uniform energy distribution, indicating natural speech production. Such characteristics serve as a baseline for distinguishing authentic speech during preprocessing and feature extraction stages.

The Fake Audio WAV Diagram presents examples of artificially synthesized speech produced through advanced voice transformation mechanisms. While such fabricated audio signals can sound highly similar to natural human speech, their waveform representations frequently reveal distinctive characteristics, including overly smooth signal flow, recurring structural patterns, or minor signal irregularities. These anomalies arise from automated speech synthesis and transformation processes and serve as valuable cues that enable analytical systems to separate manipulated voice signals from authentic recordings during examination.

6. CONCLUSION

This project focuses on addressing an emerging threat to modern information systems caused by the misuse of artificially generated voice content. As automated speech synthesis and voice manipulation technologies advance and become more accessible, they introduce serious concerns related to authenticity, identity protection, and secure communication. To mitigate these risks, the proposed work seeks to design an intelligent audio verification framework that leverages convolution-based neural architectures combined with advanced acoustic representation methods derived from time–frequency and cepstral analysis.

The objective of this initiative is to construct an efficient, adaptable, and accessible detection mechanism that can reliably differentiate authentic speech signals from manipulated audio samples. Implementation will be carried out in Python, utilizing publicly available datasets and open- source development tools to ensure reproducibility and extensibility. Beyond immediate system development, this effort aims to provide a reusable

foundation for continued exploration and real-world adoption.

By pursuing this research, the project contributes to ongoing advancements in digital investigation, media verification, and responsible artificial intelligence practices. Ultimately, it supports stakeholders—including individuals, institutions, and public bodies—in preserving credibility and accountability within rapidly evolving intelligent communication environments. This proposal represents an initial step toward delivering a functional and effective countermeasure against the growing misuse of synthetic voice technologies.

7. REFERENCES

[1] Tanveer Gujjar, M. U., Munir, K., Amjad, M., Rehman, A. U., & Bermak, A. (2024). *Unmasking the Fake: Machine Learning Approach for Deepfake Voice Detection*. IEEE Access, Accepted December 18, 2024. DOI: 10.1109/ACCESS.2024.3521026.

[2] Passos, L. A., Jodas, D., Costa, K. A. P., Souza Júnior, L. A., Rodrigues, D., Del Ser, J., Camacho, D., & Papa, J. P. (2024). *A Review of Deep Learning-based Approaches for Deepfake Content Detection*. arXiv preprint arXiv:2202.06095v3, February 15, 2024.

[3] Ganavi M., Shashank R. R., Varun S., Salanke R. S., & Navale V. S. (2025). *AudioVeritas: A Machine Learning Model to Detect Deepfake Audio*. International Journal of Research in Advent Technology (IJRASET). DOI: 10.22214/ijraset.2025.66025.

[4] F. G., K. S., M. M., and T. N., "Deepfake Audio Detection Model Based on Mel Spectrogram Using Convolutional Neural Network," *Department of Computer Science and Engineering, Adhiyamaan College of Engineering, Hosur, India*.

[5] Waseem, Saima, Syed R. Abu-Bakar, Bilal Ashfaq Ahmed, Zaid Omar, Taiseer Abdalla Elfadil Eisa, and Mhassen Elnour Elneel Dalam. "DeepFake on Face and Expression Swap: A Review." IEEE Access (2023).

[6] Abbasi, Ahmed, Abdul Rehman Rehman Javed, Amanullah Yasin, Zunera Jalil, Natalia Kryvinska, and Usman Tariq. "A large-scale benchmark dataset for anomaly detection and rare event classification for audio forensics." IEEE Access 10 (2022): 38885-38894.

[7] V. Phani and Krishna Deep, "Fake detection using LSTM and RESNEXT", Journal of Engineering Sciences, vol. 13, no. 07, July 2022, ISSN 0377–9254. [4] Pramod
Dhamdhere, “Semantic trademark retrieval system based on conceptual similarity of text with leveraging histogram computation for images to reduce trademark infringement”, Webology (ISSN: 1735- 188X), Volume 18, Number 5,
2021.

[8] J. Khochare, C. Joshi, B. Yenarkar, S. Suratkar, F. Kazi, A deep learning framework for audio deepfake detection, Arabian Journal for Science and Engineering (2021) 1–12.

[6] D. Cozzolino, M. Nießner and L. Verdoliva, "Audio- visual person-of-interest deepfake detection", 2022.

[9] angyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, et al., "Add 2022: the first audio deep synthesis detection challenge", 2022 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), 2022.

[10] Y. Gao, T. Vuong, M. Elyasi, G. Bharaj and R. Singh, "Generalized Spoofing Detection Inspired from Audio Generation Artifacts", 2021.