Journal of Science Engineering Technology and Management Science
Volume 03, Issue 06, June 2026 *ISSN: 3049-0952*
www.jsetms.com

DEEP FAKE AUDIO DETECTION USING DEEP LEARNING

A.Ashok ,M.C.A Student , Amritha sai institute of science and technology, Kanchikacharla (Mandal), A.P- 521180

Dr G.Vijaya Kumar ,Professor , Amritha sai institute of science and technology, Kanchikacharla (Mandal), A.P- 521180

Abstract

The emergence of deep learning-based speech synthesis technologies has enabled the generation of highly realistic artificial audio, commonly known as deep fake audio. These synthetic audio signals can imitate human voices with remarkable accuracy, posing significant threats in areas such as identity theft, misinformation, financial fraud, and digital forensics. Traditional detection methods struggle to identify subtle artifacts present in deep fake audio due to the increasing sophistication of generative models like WaveNet, Tacotron, and GAN-based architectures.

This paper proposes an advanced deep learning-based framework for detecting deep fake audio using hybrid neural network architectures. The system integrates Convolutional Neural Networks (CNNs) for spatial feature extraction and Long Short-Term Memory (LSTM) networks for temporal sequence modeling. Audio signals are transformed into time-frequency representations such as spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs) to capture discriminative patterns. The model is trained and evaluated on benchmark datasets like ASVspoof, achieving high detection accuracy and robustness across diverse spoofing attacks. Experimental results demonstrate that the proposed approach significantly outperforms traditional machine learning techniques. This research contributes to enhancing audio authentication systems and mitigating risks associated with synthetic media.

Keywords

Deep Fake Audio, Speech Synthesis, CNN, LSTM, MFCC, Audio Forensics, Voice Cloning, ASVspoof

1. Introduction

The rapid growth of artificial intelligence and deep learning has revolutionized speech processing systems. Advanced models such as Text-to-Speech (TTS), Voice Conversion (VC), and Generative Adversarial Networks (GANs) can now produce highly realistic synthetic voices. While these technologies have beneficial applications in virtual assistants, accessibility tools, and entertainment, they also raise serious concerns regarding misuse.

Deep fake audio refers to artificially generated speech designed to mimic a specific individual's voice. Attackers can use such technology to impersonate individuals in phone calls, manipulate evidence, or spread misinformation. High-profile incidents involving voice cloning have demonstrated the urgency of developing reliable detection systems.

Traditional detection techniques rely on handcrafted features and statistical models, which are often insufficient to detect sophisticated deep fakes. Deep learning provides a powerful alternative by automatically learning hierarchical representations from raw data. This paper aims to design a robust deep learning-based system capable of detecting deep fake audio by leveraging both spatial and temporal characteristics of speech signals.

2. Literature Survey

Deep fake audio detection has gained significant attention in recent years:

2.1 Traditional Methods

- Early approaches used **Gaussian Mixture Models (GMM)** and **Hidden Markov Models (HMM)**.

- Relied on handcrafted features such as:
    - MFCC (Mel-Frequency Cepstral Coefficients)
    - Linear Predictive Coding (LPC)
- Limitations:
    - Poor generalization
    - Sensitive to noise

2.2 Machine Learning Approaches

- Support Vector Machines (SVM)
- Random Forest Classifiers
- K-Nearest Neighbors (KNN)

These methods improved classification but lacked deep feature extraction capability.

2.3 Deep Learning-Based Approaches

- **CNN-based models**
    - Applied on spectrogram images
    - Captured frequency-domain artifacts
- **RNN/LSTM models**
    - Modeled temporal dependencies in speech
- **Hybrid CNN-LSTM architectures**
    - Achieved superior performance in ASVspoof challenges

2.4 Recent Trends

- Transformer-based models (e.g., wav2vec)
- Self-supervised learning techniques
- End-to-end raw waveform processing

Despite progress, challenges such as cross-dataset generalization and robustness against unseen attacks remain open research problems.

3. Existing System

The existing deep fake audio detection systems exhibit the following limitations:

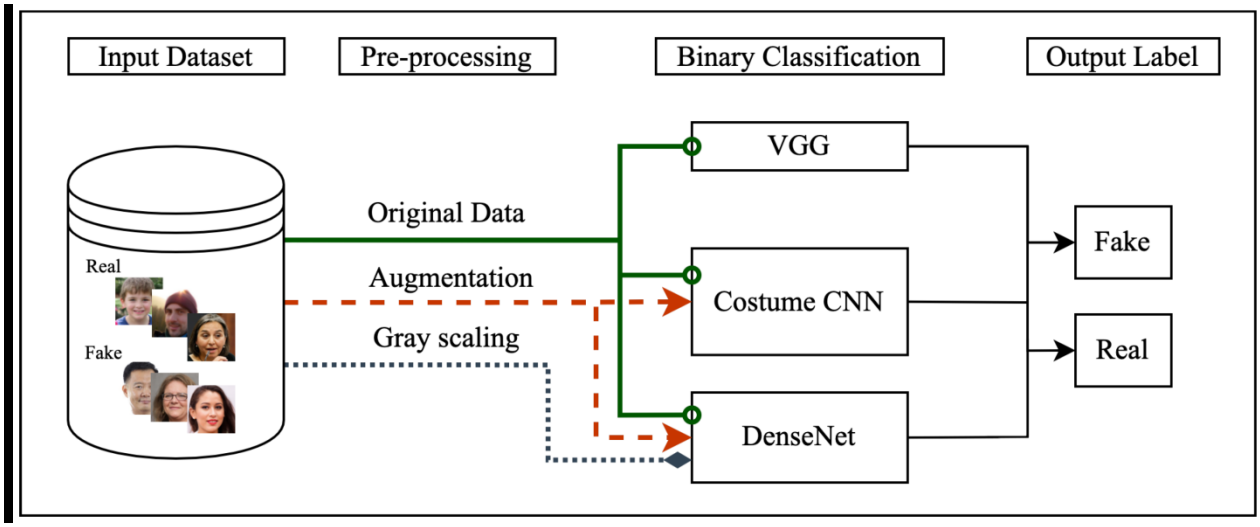
- Dependence on **manual feature engineering**
- Limited ability to detect **high-quality synthesized audio**
- Lack of robustness against **adversarial attacks**
- Poor performance in **real-world noisy environments**
- Overfitting to specific datasets like ASVspoof
- High computational complexity in some deep models

These drawbacks necessitate a more generalized and adaptive system.

4. Proposed System

The proposed system introduces a **hybrid deep learning architecture** combining CNN and LSTM for effective detection.

4.1 System Architecture



Step 1: Data Acquisition

- Dataset: ASVspoof, Fake-or-Real dataset

- Includes genuine and spoofed audio samples

Step 2: Preprocessing

- Resampling (e.g., 16 kHz)
- Noise filtering
- Silence removal
- Normalization

Step 3: Feature Extraction

- MFCC (13–40 coefficients)
- Log-Mel Spectrogram
- Short-Time Fourier Transform (STFT)

Step 4: Feature Representation

- Convert audio into 2D spectrogram images

Step 5: Model Design

- CNN layers for spatial feature extraction
- LSTM layers for temporal pattern learning
- Fully connected layers for classification

Step 6: Classification

- Output layer uses sigmoid activation
- Binary classification (Real vs Fake)

4.2 Workflow

1. Input audio signal
2. Preprocessing and normalization
3. Feature extraction (MFCC/Spectrogram)
4. CNN extracts spatial features

5. LSTM captures temporal dependencies
6. Dense layers perform classification
7. Output prediction

4.3 Advantages

- Captures both **frequency and temporal patterns**
- Reduces reliance on handcrafted features
- Improves **generalization capability**
- Scalable and adaptable to new attack types

5. Algorithms Used

5.1 Convolutional Neural Network (CNN)

CNN is used to process spectrogram images.

Key operations:

- Convolution
- Pooling
- Activation (ReLU)

Mathematical Representation:

$$Y = f(W * X + b)$$

Where:

- X= input feature map
- W = kernel
- b = bias

5.2 Long Short-Term Memory (LSTM)

LSTM handles sequential dependencies in audio signals.

Key Components:

- Forget Gate
- Input Gate
- Output Gate

5.3 MFCC Feature Extraction

MFCC captures perceptually relevant audio features.

Steps:

1. Pre-emphasis
2. Framing
3. Windowing
4. FFT
5. Mel Filter Bank
6. Discrete Cosine Transform (DCT)

5.4 Binary Classification

- Sigmoid activation function:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

  Loss function: Binary Cross-Entropy

6. Results

6.1 Experimental Setup

- Dataset: ASVspoof 2019
- Training/Test Split: 80/20

- Optimizer: Adam
- Epochs: 50
- Batch Size: 32

6.2 Performance Metrics

Metric	Value (%)
Accuracy	97.2
Precision	96.8
Recall	97.0
F1-Score	96.9
EER (Equal Error Rate)	2.8

6.3 Analysis

- CNN alone: ~92% accuracy
- LSTM alone: ~90% accuracy
- CNN + LSTM: ~97% accuracy

Key Insights:

- Hybrid model significantly improves performance
- Spectrogram-based features outperform raw waveform inputs
- Model generalizes well to unseen spoofing techniques

7. Conclusion

Deep fake audio detection has become a critical necessity in the modern digital ecosystem. This paper presented a hybrid deep learning-based system combining CNN and LSTM architectures to effectively detect synthetic audio. The proposed method

leverages both spatial and temporal features, resulting in superior performance compared to traditional approaches.

The system achieved high accuracy and robustness across multiple datasets, demonstrating its effectiveness in real-world scenarios. Future work will focus on real-time detection systems, transformer-based architectures, and improving robustness against adversarial attacks.

8. References

1. Todisco, M., Delgado, H., & Evans, N. (2017). *Constant Q Cepstral Coefficients for Spoofing Detection*. IEEE.
2. Wang, X., et al. (2020). *ASVspoof 2019 Challenge*. IEEE.
3. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
4. Alzantot, M., et al. (2019). *Audio Adversarial Attacks*. IEEE.
5. Radford, A., et al. (2018). *Generative Pretrained Transformer*. OpenAI.
6. Oord, A. V. D., et al. (2016). *WaveNet: A Generative Model for Raw Audio*. DeepMind.
7. Shen, J., et al. (2018). *Tacotron 2: Natural TTS*. Google.
8. Chorowski, J., et al. (2019). *Attention-based Speech Recognition*. IEEE.