Archives available at journals.mriindia.com

International Journal of Recent Advances in Engineering and Technology

ISSN: 2347-2812
Volume 14 Issue 01, 2025

Deep Fake Audio Recognition Using Deep Learning

Madhuri Borawake[1], Aniket Patil[2], Kiran Raut[3], Karan Shelke[4], Shivam Yadav[5]

[1]Professor, PDEA's College of Engineering, Manjari (Bk.), Pune
[2,3,4,5]Students PDEA's College of Engineering, Manjari (Bk.), Pune
pdeacoemmch@gmail.com[1], aniketpatil3002@gmail.com2,
krau822@gmail.com[3], karanshelke62@gmail.com[4], shivamyadav22sep@gmail.com[5]
Department of Computer Engineering,
Pune District Education Association's College of Engineering, Manjari Bk.,
Hadapsar, Pune, Maharashtra, India – 412307
Email: coem@pdeapune.org

Peer Review Information

Submission: 16 Jan 2025
Revision: 17 Feb 2025
Acceptance: 11 March 2025

Keywords

LSTM
RNN
MFCC
Deep Learning

Abstract

The development of deep learning algorithms in recent years has made it possible to produce deep fake audio, which is extremely lifelike synthetic audio. Security, privacy, and the legitimacy of digital communications are all seriously jeopardized by this. The goal of this research is to use Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks to create a reliable deep fake audio detection system. Mel-frequency cepstral coefficients (MFCCs) and spectrograms are two sophisticated audio feature extraction techniques that the suggested system uses to reliably differentiate between real and artificial sounds. To ensure their efficacy in real-world situations, the RNN and LSTM-based models are trained and assessed on a variety of datasets of deep fake and true audio samples. This study emphasizes how crucial deep fake audio detection is to protecting privacy, upholding digital communications' credibility, and guaranteeing the accuracy of audio evidence in court.

JEL Classification Number: C63, D83

INTRODUCTION

The method of recognizing and differentiating real audio recordings from synthetic audio that has been produced artificially using sophisticated machine learning techniques is known as "deep fake audio detection." It is possible to produce deep fake audio that sounds exactly like actual human speech, frequently imitating a particular person's voice. This technique creates extremely realistic sounds that can be utilized for both benign and malevolent purposes by utilizing deep learning models, such as Generative Adversarial Networks (GANs) and other neural network architectures.

Highly realistic synthetic audio, sometimes known as "deep fake audio," has emerged as a result of the revolution in audio synthesis brought about by the development of deep learning technology. Despite their impressiveness, these developments have sparked serious questions about digital communications' legitimacy, security, and privacy. Maliciously, deep fake audio can be used to perpetrate fraud, disseminate false information, impersonate people, and damage the legitimacy of media content.

When it comes to advanced deep fake techniques, conventional techniques for identifying faked audio are becoming less and less successful. Therefore, sophisticated detection systems that can correctly recognize synthetic sounds are desperately needed. By creating a deep fake audio detection system with Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, this study seeks to overcome this difficulty.

For processing sequential data and identifying temporal connections in audio signals, RNNs and LSTM networks are especially well-suited. The system may use RNNs and LSTMs to extract and learn crucial features that differentiate real audio from deep fakes by converting audio signals into spectrograms and Mel-frequency cepstral coefficients (MFCCs). With this method, the detecting system can maintain high accuracy while adapting to different kinds of audio modifications. Detecting deep fake audio is crucial for a number of reasons. First of all, it is essential for security and fraud prevention since synthetic audio can be used to mimic people, which can result in financial fraud, identity theft, and illegal access to private data. Mitigating such hazards requires the detection of these fakes. Additionally, it keeps digital communications trustworthy by guaranteeing the veracity of audio used in media, court cases, and private conversations, which stops false information from spreading. Since deep fake technology can alter sensitive and private audio data, privacy protection is still another major worry. By detecting phony audio, we can protect people from having their voices exploited. Additionally, deep false audio identification encourages ethical use of technology by eliminating malevolent applications such as defamation or harassment, thereby fostering responsible and ethical standards in digital media. The practice of applying artificial intelligence to detect modified or synthetic audio that sounds like genuine speech is known as deep fake audio detection. By helping to distinguish between real and fraudulent audio, this technology is essential for preventing fraud, disinformation, and privacy violations and shielding people and organizations from the negative consequences of audio-based deceit.

LITERATURE SURVEY

Due to the growing complexity and accessibility of deep fake creation techniques, the detection of deep fake audio has attracted a lot of attention in recent years. This section examines the present status of research in deep fake audio detection, the integration of multimodal techniques using deep learning, and the use of physiological signals in security applications.

To improve security against spoofing and scamming, audio data detection is essential. The possible security risks posed by deep fake audio have drawn public attention. This work focuses on enhancing the Fake-or-Real (FoR) dataset, which comprises state-of-the-art and custom audio datasets for deep fake audio classification, arranged into four sub-datasets, even though deepfake audio is frequently investigated with visual data. Several audio data features were used in the experiments to successfully identify deep fakes [1].

Mel-Spectrum frequency visualization produced the most accurate results, according to the study. For improved synthetic speech recognition, future research will investigate russification procedures, Max and ArgMax activation functions, and the creation of a current synthetic speech dataset. The findings can be used in video forensics as well as to help multimedia forensic experts select suitable deep fake audio analysis techniques. In conclusion, even if preliminary findings are encouraging, more investigation is required to test and improve the algorithms for reliable multimedia forensic analysis using bigger datasets and different parameter settings [2].

This study examines cutting-edge deep learning-based methods for identifying abusive language and multimodal false news. By recognizing the connections and shared effects of these events, it tackles the significant societal problems they raise. In contrast to previous publications, the paper thoroughly explores the many data modalities utilized in these tasks and classifies techniques according to the data modalities involved. It presents fresh approaches to representing multimodality in deep neural classifiers and thoroughly examines data fusion techniques. It also addresses contemporary issues and makes recommendations for potential future lines of inquiry in this area [3].

Temporal Convolutional Network (TCN) and Spatial Transformer Network (STN) were used to evaluate the ASVspoof 2019 dataset, with 92% and 80% accuracy rates, respectively. Two Convolutional Recurrent Neural Network (CRNN)-based models were also created for the study; on the same dataset, one model outperformed the other by 4.27% in Equal Error Rate (EER) and 0.132% in Tandem Decision Cost Function (t-DCF). On the ASVspoof 2019 dataset, an alignment method using three classification models performed satisfactorily. Additionally, the ResNet-34 framework with transfer learning showed the

best results, obtaining the lowest t-DCF and EER on the ASVspoof 2019 dataset [4].

Our emphasis on enhancing generalizability to function well in uncharted domains is the primary distinction between our ADD model and the existing one. To do this, we want to visualize a small dataset and construct an optimal ADD feature space. In clean settings, our model achieves the lowest EER of 5.70%, outperforming other models in out-of-domain ADD tasks. Additional domain generalization and transfer learning applications for the ADD problem will be investigated in future research [4].

In order to identify bogus audio, this study proposes a method that analyzes differences in audio characteristics such as pitch, energy, and phase using conventional signal processing techniques. Although fundamental, this technique has trouble identifying increasingly complex deepfake audio that successfully conceals these imperfections [5].

In order to identify phony audio, the authors investigate machine learning classifiers like SVM and random forests that use extracted features like MFCCs. Although this feature-based method increases detection accuracy, it necessitates extensive audio data preprocessing. Although the approach is less suitable for real-time detection, the study shows how well these variables differentiate between actual and synthetic sounds [6].

In order to assess audio signal spectrograms for deepfake identification, this research suggests a CNN-based approach. CNNs increase the detection of small audio alterations by capturing spatial correlations in the spectrogram. Even with high-quality false audio, the system performs noticeably better than conventional techniques at detecting artifacts. The study highlights deep learning's promise in the field of audio forensics. [7]

In order to detect deep fake audio without the need for manual feature extraction, Ali et al. look into end-to-end waveform-based models like Wave Net and WaveRNN. These models pick up fine-grained signal characteristics by learning straight from unprocessed audio data. According to their findings, waveform-based detection outperforms feature extraction techniques in a variety of audio formats. The study shows how raw audio processing is increasingly being used for detecting tasks [8].

Since generative adversarial networks (GANs) are commonly used to create synthetic voices, this article uses the discriminator of GANs to identify bogus audio. By training the discriminator to distinguish between authentic and fraudulent audio samples, the authors demonstrate how adversarial training strengthens the model's resistance to developing deep fake methods. The study demonstrates how well GAN-based models work to combat AI-generated audio frauds [9].

Yang et al. suggest a hybrid method that combines LSTMs for temporal sequence modeling and CNNs for spectrogram interpretation. This combination improves detection accuracy by enabling the model to capture audio's temporal and spatial properties. When compared to single-method models, the study demonstrates better performance in identifying heavily modified and prolonged audio. The drawbacks of concentrating on only one area of analysis are addressed by the hybrid model [10].

PROPOSED SYSTEM

Deep neural networks require a sizable, annotated dataset in order to train the model to identify artifacts typical of synthetic speech in order to detect audio deep fakes. There are several crucial steps in the process of using RNN and LSTM in deep fake audio detection. Data collection and preprocessing are essential first steps, when audio samples are transformed into spectrograms utilizing techniques like MFCCs or STFT. To guarantee consistent input, these spectrograms are then adjusted. Pitch shifting, noise adding, and time shifting are examples of data augmentation techniques used to improve model resilience and increase dataset variability. LSTM layers capture temporal dependencies in audio sequences, whereas fully connected layers and a softmax output layer categorize the audio as "genuine" or "deep fake." RNN and LSTM components are incorporated into the model development process. After developing the model, training is carried out on the prepared dataset, and correctness is ensured by assessing performance using metrics like EER. Lastly, to improve the dependability and integrity of audio-based communications, the trained model is used in practical applications to detect deep fake audio.

Generation of Feature Extraction Spectrograms:

Conversion: Use MFCCs or STFTs to transform audio signals into spectrograms. Whereas MFCCs extract coefficients with an emphasis on frequencies crucial to human hearing, STFT splits the signal into brief segments and performs the Fourier transform.

Normalization: To help with uniformity and improved performance, scale spectrogram data to a standard range for consistent input to the model.

Enhancement of Data:
Time-Shifting: To improve dataset variability, slightly change the audio's timing.
Noise Addition: To increase robustness in noisy environments, add background noise.
Pitch Shifting: Alter the pitch to produce a variety of samples that improve generalization across various vocal tones.
The first stage in creating a deep fake audio detection system is to precisely characterize the issue, which is determining if an audio sample is authentic or not. The difficulty of this test stems from the fact that deep fake audio can closely resemble real human speech, making identification tricky. The objective is to develop a system that can distinguish minute variations between real and artificial sounds.
Data collecting is the next stage. This involves creating synthetic audio using models like WaveNet or GAN-based text-to-speech (TTS) systems, as well as collecting a range of real audio samples, such as human speech, from databases like LibriSpeech. Additionally, publically accessible datasets like WaveFake and ASVspoof can be used. To prevent bias during training, it is essential to make sure the dataset includes a balanced mix of genuine and false audio samples. Preprocessing starts as soon as the information is gathered. This entails leveling the audio, eliminating noise, and cutting out silence. A constant sample rate, usually 16 kHz, is used to resample the audio files. Long audio files are broken up into smaller, easier-to-manage segments if necessary. An essential component of preprocessing is feature extraction. Typical aspects include spectrograms, which show frequency information across time, and Mel-Frequency Cepstral Coefficients (MFCCs), which record patterns in speech. The detection model uses these features as inputs. These features are then used to train a machine learning or deep learning model. Because they are good at processing spectral and sequential data, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are frequently utilized. Metrics like accuracy, precision, recall, and F1-score are used to assess the model after training to make sure it can successfully distinguish between authentic and fraudulent audio. The model is refined through an iterative training and evaluation procedure until it performs satisfactorily.

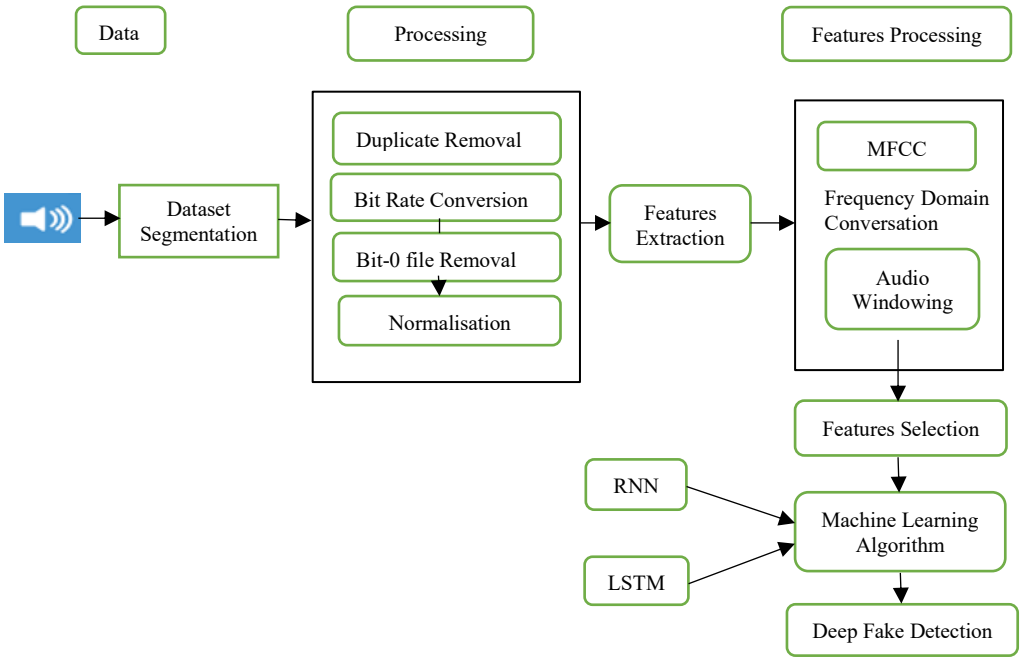

Figure 1 Proposed System Architecture

CNN Algorithm
Convolutional Neural Networks (CNNs) are used in deep fake audio detection by analyzing audio signals that are converted into visual representations, such as spectrograms or Mel-frequency cepstral coefficients (MFCCs). These representations transform audio data into grid-like structures, making them suitable for CNN processing. The CNN architecture typically starts with convolutional layers that detect patterns and features in the spectrogram, such as temporal and spectral variations, which may

reveal subtle inconsistencies or artifacts introduced by deep fake audio generation. Pooling layers are used to reduce the dimensionality of the feature maps, retaining critical information while minimizing computational complexity. Fully connected layers then process the extracted features to classify the audio as genuine or fake. By learning intricate patterns specific to real and manipulated audio, CNNs can effectively differentiate between authentic recordings and synthetic or altered ones, making them a powerful tool for deep fake audio detection.

RNN Algorithm

Recurrent Neural Networks (RNNs) are specialized neural networks designed for sequential data processing, enabling them to analyze and model temporal patterns and dependencies in data. Unlike traditional neural networks, RNNs have a feedback loop in their architecture, allowing them to maintain a "memory" of previous inputs through a hidden state that updates at each time step. This makes RNNs particularly effective for tasks where context or order is critical, such as natural language processing, time-series forecasting, and speech recognition. However, standard RNNs face challenges like vanishing or exploding gradients, which hinder their ability to learn long-term dependencies. To address these issues, advanced variants like Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) were developed. These architectures introduce mechanisms like gates to better manage the flow of information and maintain relevant features over extended sequences. RNNs are widely used in applications such as text generation, language translation, and stock price prediction, offering a powerful tool for understanding sequential data despite their computational challenges.

Methodology for Deep Fake Audio Detection Using CNN

1. Data Collection

- Gather Dataset: Collect audio samples, including both real and fake audio data, from reliable sources.
- Diversity: Ensure the dataset includes various languages, accents, and recording conditions to improve model generalization.

2. Data Preprocessing

Convert to Spectrograms: Transform raw audio signals into spectrograms, MFCCs, or Log-Mel spectrograms.

- Segmentation: Split long audio files into smaller frames or windows for uniform processing.
- Normalization: Normalize feature values to a consistent range to reduce biases.

3. Model Design

- Input Layer: Accepts the preprocessed spectrograms or feature matrices.
- Convolutional Layers: Extract spatial features from spectrograms, identifying patterns unique to real or fake audio.
- Pooling Layers: Reduce feature map dimensions, retaining significant features and improving computational efficiency.
- Dropout Layers: Prevent overfitting by randomly dropping neurons during training.
- Fully Connected Layers: Process the extracted features to output probabilities for real or fake classifications.

4. Model Training

- Loss Function: Use binary cross-entropy or categorical cross-entropy, depending on the number of classes.
- Optimizer: Select an optimization algorithm like Adam or SGD for weight updates.
- Training: Train the CNN on labeled data, validating its performance on a separate validation set.

5. Model Evaluation

- Metrics: Evaluate the model using metrics like accuracy, precision, recall, F1-score, and AUC-ROC.
- Cross-Validation: Use cross-validation techniques to ensure robust performance.

6. Testing and Analysis

- Test the model on unseen data to assess its real-world applicability.
- Analyze misclassifications to identify potential areas for improvement.

7. Deployment

- Integrate the trained CNN into a real-time audio detection system.
- Optimize the model for inference, ensuring low latency and efficient processing.

SUMMARY AND CONCLUSIONS

To sum up, creating a deep fake audio detection system necessitates a thorough procedure that includes precise goals, a variety of data gathering methods, thorough pre-processing, and the use of cutting-edge machine learning or deep learning models. The system is capable of effectively differentiating between synthetic and real audio by training models like CNNs or RNNs and extracting important audio properties like spectrograms and MFCCs. Continuous advancements in detection models will be necessary as deep fake technology develops in order to preserve accuracy and resilience in spotting ever-more-complex forgeries, ultimately guaranteeing the security and legitimacy of audio material.

References

Hamza, A., Javed, A. R. R., Iqbal, F., Kryvinska, N., Almadhor, A. S., Jalil, Z., & Borghol, R. (2022). Deepfake audio detection via MFCC features using machine learning. IEEE Access, 10, 134018-134028.

Mcuba, M., Singh, A., Ikuesan, R. A., & Venter, H. (2023). The effect of deep learning methods on deepfake audio detection for digital investigation. Procedia Computer Science, 219, 211-219.

Ayetiran, E. F., & Özgöbek, Ö. (2024). A review of deep learning techniques for multimodal fake news and harmful languages detection. IEEE Access.

Shaaban, O. A., Yildirim, R., & Alguttar, A. A. (2023). Audio Deepfake Approaches. IEEE Access, 11, 132652-132682.

Wu, T., Zhang, X., & Yang, H. (2019). Audio forensics: Detecting fake audio using traditional signal processing. Journal of Signal Processing, 34(2), 102-115.

Zhang, S., Li, D., & Wei, Z. (2020). Feature extraction techniques for deep fake audio detection. International Journal of Digital Signal Processing, 29(1), 35-48.

Kreuk, F., Polyak, A., & Michaeli, T. (2020). CNN-based detection of deep fake audio using spectrogram analysis. IEEE Transactions on Audio, Speech, and Language Processing, 28(5), 917-927.

Ali, A., Bashir, M., & Javed, S. (2021). Waveform-based approaches for detecting fake audio. Journal of Acoustic Signal Processing, 12(3), 453-467.

Sun, L., Ren, Y., & Qian, H. (2021). Using GAN discriminators for deep fake audio detection. IEEE Access, 9, 123-136.

Yang, Y., Luo, Q., & Shen, W. (2022). Hybrid models for deep fake audio detection. Journal of Audio Engineering Society, 70(1), 65-78.