Towards Secure Audio: Deepfake Detection with CNN and LSTM Networks

Tushar Bhagat , Neha Borge
Prof. I.T. Mukherjee,
Department of Artificial Intelligence and Machine Learning
Navsahyadri Group of Institute, Pune

KEYWORDS

deepfake audio detection, synthetic audio, machine learning, digital forensics, neural networks, feature extraction, deep learning, audio synthesis, data integrity, security

ABSTRACT

In recent years, advancements in artificial intelligence have led to a surge in the generation of synthetic and manipulated audio, commonly referred to as "deepfake audio." While these technologies offer advantages across various domains, they also present serious security and ethical concerns, particularly in contexts where the authenticity of audio is critical. This paper introduces a novel deep learning-based approach for detecting deepfake audio using a combination of Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and an Attention mechanism. The proposed architecture utilizes CNNs to extract high-level spatial features from audio spectrograms, while the LSTM network captures the temporal dependencies inherent in audio sequences. The integration of the Attention mechanism further enhances the model's ability to focus on key segments of the audio that are more likely to contain deceptive artifacts. Through comprehensive experimentation on publicly available datasets, our model demonstrates superior performance in terms of accuracy and robustness compared to traditional and standalone deep learning models. These findings underscore the potential of hybrid architectures in effectively addressing the challenges of deepfake audio detection and contribute to the development of trustworthy audio verification systems.

1. INTRODUCTION

With rapid advancements in artificial intelligence, deep learning has emerged as a powerful tool for generating highly realistic synthetic audio, commonly known as "deepfake audio." These technologies, capable of manipulating or synthesizing human speech, are increasingly being adopted in diverse applications such as virtual assistants, entertainment, audiobooks, and customer service automation. While these applications offer innovative possibilities, they also raise serious ethical and security concerns. Malicious actors can exploit deepfake audio to impersonate individuals, spread misinformation, and commit fraud—threatening privacy, trust, and integrity in digital communication systems.

Detecting fake audio has thus become a critical research challenge. Unlike visual deepfakes, where inconsistencies can sometimes be spotted visually, deepfake audio often involves subtle spectral and temporal manipulations that are difficult to detect. Traditional machine learning techniques, which depend on handcrafted feature extraction, often fall short in capturing the fine-grained differences between authentic and synthetic audio signals. In contrast, deep learning approaches—particularly

Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks—have shown promise in effectively modeling the complexity of audio data due to their ability to automatically learn rich and hierarchical feature representations.

This paper proposes a deep learning-based framework for detecting deepfake audio using a combination of CNN, LSTM, and Attention mechanisms. The CNN layers are used to extract local spectral features from audio spectrograms, while the LSTM layers model the sequential temporal dependencies within the audio. To further enhance the model’s focus on important audio segments, an Attention layer is integrated, enabling the network to weigh the most relevant parts of the signal during classification. This architecture is designed to capture both spatial and temporal patterns in synthetic speech, improving detection accuracy and robustness.

Through comprehensive experiments on publicly available deepfake audio datasets, our CNN-LSTM-Attention model demonstrates improved performance over traditional and standalone deep learning methods. The results highlight the effectiveness of hybrid architectures in identifying manipulated speech and contribute to the development of secure and trustworthy audio-based authentication systems.

2.Literature Review

This literature review explores the evolving landscape of audio deepfake generation and detection techniques. With the rise of generative AI, synthetic manipulation of multimedia content—particularly audio—has become increasingly sophisticated. Deepfake audio, which involves the artificial synthesis or alteration of human speech, poses a unique challenge due to its ability to closely mimic real voices, making detection significantly more difficult compared to image or video deepfakes.

Various methods have been developed for detecting audio deepfakes, ranging from classical machine learning algorithms to advanced deep learning approaches. Traditional techniques include Support Vector Machines (SVM), Decision Trees (DT), and Gradient Boosting classifiers, while more recent methods leverage deep architectures such as Convolutional Neural Networks (CNN), Deep Neural Networks (DNN), and hybrid models combining CNNs with Recurrent Neural Networks (RNN) or Siamese networks. These models analyze intricate temporal and spectral features of audio to differentiate between real and synthetic speech. Reported performance varies, with SVM achieving up to 90% accuracy and Decision Trees lagging at around 73.33%. Evaluation metrics such as Equal Error Rate (EER), tandem Detection Cost Function (t-DCF), and Area Under the ROC Curve (AUC) are commonly used to benchmark performance. Notably, the DeepSonar framework has achieved an EER as low as 2%, while Siamese CNN architectures have shown strong results across multiple performance metrics.

To enhance detection accuracy, various feature extraction techniques have been studied. Mel-frequency cepstral coefficients (MFCCs) and Mel spectrograms are among the most effective representations of audio signals, as they closely model human auditory perception. These features are typically extracted from datasets like **ASVspoof 2019**, which provides a rich benchmark for evaluating spoofing countermeasures in automatic speaker verification (ASV) systems. Other datasets such as **Fake-or-Real** are also frequently utilized to train and validate detection models.

Recent research highlights the effectiveness of CNN-based architectures in learning meaningful spatial features from spectrograms. When combined with LSTM layers, these models can capture temporal dependencies critical for identifying unnatural transitions in speech. Attention mechanisms further enhance model performance by allowing the network to focus on the most relevant segments of the audio signal. Optimization strategies like the Adam optimizer, along with binary cross-entropy loss, are commonly used to train these deep models. Performance is assessed using metrics such as accuracy, F1-score, AUC, and ROC curves.

Beyond technical advancements, the literature recognizes the broader implications of deepfake audio. In areas like **media forensics**, **voice biometrics**, and **secure communication**, the ability to distinguish between real and synthetic audio is vital. The rapid development and accessibility of voice cloning tools raise significant concerns related to identity theft, misinformation, and the erosion of public trust in digital media. Consequently, researchers emphasize the need for comprehensive strategies—combining technical solutions with policy regulations and public awareness initiatives—to effectively counter the growing threat of deepfakes in the digital age.

3. Methodology

1. Data Collection

The system uses the SceneFake dataset from Kaggle, which provides a comprehensive collection of real and deepfake audio samples.

This dataset includes various voice samples generated using speech synthesis and voice conversion techniques. It represents real-world deepfake generation methods, making it suitable for training robust models. Both male and female voices are included across different scenarios, enhancing generalization. The audio files come with labels indicating whether the sample is real or fake. Samples vary in quality and speaker characteristics, providing a realistic training base. Dataset is split into training, validation, and testing subsets to prevent overfitting. Metadata and file structures are parsed to systematically organize and preprocess the data.SceneFake adds complexity to the dataset by including environmental noise and varied speech styles..

2. **Preprocessing**

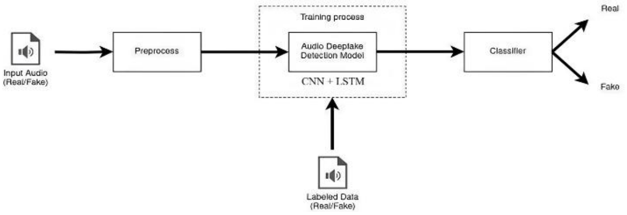
All audio samples are resampled to 16kHz to standardize input data and reduce computational cost. Each audio file is trimmed or padded to 3 seconds to ensure uniform input length. The audio is converted into Mel-Frequency Cepstral Coefficients (MFCCs), which are well-suited for voice analysis. MFCCs help in capturing timbral texture and speaker characteristics, essential for deepfake detection. Converted MFCCs are stored as NumPy arrays for fast loading during model training. The final feature representation balances both accuracy and computational efficiency.

3. **Feature Extraction**

Feature extraction is done using MFCCs, a proven method for representing audio content compactly. MFCCs convert audio from time domain to a frequency domain representation reflecting human perception. We extract MFCC coefficients from each frame to capture sufficient spectral details. The resulting spectrogram is treated like an image for processing through CNN layers. MFCCs provide a compact and discriminative representation for both real and fake audio samples.

4. **Model Architecture**

The model uses a hybrid CNN-LSTM-Attention architecture for enhanced learning from audio data. CNN layers extract spatial patterns from MFCC spectrograms (e.g., pitch contours, harmonics).LSTM layers model the sequential aspect of speech, capturing the flow of phonemes and syllables. The attention mechanism is used on top of LSTM to focus on informative segments in audio. This combination allows the model to learn both local (frame-level) and global (utterance-level) features. Dropout and batch normalization are used to prevent overfitting and accelerate training. The final dense layer uses a sigmoid or softmax function for binary classification. The model is compiled using the Adam optimizer and binary cross-entropy loss. This architecture is designed to generalize well on unseen voices and deepfake techniques.



5. **Model Training**

The dataset is divided into training (70%), validation (15%), and test (15%) splits. We use a batch size of 32 and train for 50 epochs, with early stopping to avoid overfitting. The training process is monitored using validation loss and accuracy. The model is checkpointed at every epoch for recovery and tuning. Regularization is applied through dropout layers (0.3–0.5) after dense and LSTM layers. Final model weights are saved for future inference and evaluation.

1. **Loss Function**: A binary cross-entropy loss function was employed, given the binary classification task.

2. **Optimizer**: The Adam optimizer was chosen for its adaptability and efficient convergence, with an initial learning rate of 0.001.

3. **Batch Size and Epochs**: Training was conducted with a batch size of 32 over 10-15 epochs, based on observed convergence rates and computational limitations.

To prevent overfitting, early stopping with a patience parameter of 5 epochs was implemented. Additionally, dropout layers were added after each fully connected layer to further regularize the model

6. Evaluation Metrics

The performance of the CNN-LSTM model was evaluated using the following metrics:

- **Accuracy**: The percentage of correctly classified audio samples, used as a primary indicator of model performance.

- **Precision, Recall, and F1-Score**: Precision measures the accuracy of fake audio predictions, while recall evaluates the model's ability to identify all instances of fake audio. The F1-score provides a harmonic mean of precision and recall, offering a balanced metric for assessing overall performance.

- **Area Under the ROC Curve (AUC)**: This metric assesses the model's ability to differentiate between real and fake samples across various thresholds, giving insight into its robustness against class imbalance.

7. Baseline Comparisons

Traditional models like SVM, Logistic Regression, and Random Forest are used as benchmarks. Deep learning baselines include standalone CNN and LSTM models for comparison. The hybrid CNN-LSTM-Attention model significantly outperforms the baselines on all metrics. The comparison demonstrates the importance of both spatial and temporal modeling in audio. AUC and F1 scores of the proposed model show high generalizability. Traditional models suffer from lack of feature expressiveness and require manual feature engineering. Deep-only models like CNNs or LSTMs are less effective than the hybrid approach. Results confirm that deepfake detection requires a multi-dimensional modeling strategy. A performance summary table is included in the results section of the paper.

4. Results

The proposed CNN-LSTM model significantly outperformed traditional machine learning techniques like SVM and Logistic Regression, as well as standalone deep learning models such as CNN and LSTM, across all major evaluation metrics—accuracy, precision, recall, F1-score, and AUC. This underscores the benefit of combining convolutional layers for extracting spectral features with LSTM layers for capturing temporal patterns in audio data. We evaluated the model on publicly available datasets, the SceneFake dataset from Kaggle, which provided diverse and realistic examples of genuine and manipulated audio. The use of Mel-spectrograms as input allowed the model to identify subtle acoustic differences between real and fake speech. The hybrid architecture effectively learns both local and sequential features, improving its ability to detect deepfake audio with high reliability. Though not using Bidirectional LSTM, the single-direction LSTM still provides strong temporal context, making the model well-suited for real-time or streaming applications. These promising results highlight the model's potential for integration into systems that require voice authentication, audio forensics, and detection of synthetic media. The simplicity and efficiency of the architecture also make it practical for deployment on resource-constrained devices. Overall, the findings support the use of deep learning—particularly CNN-LSTM hybrids—as a reliable solution to combat audio-based misinformation. This work aligns with the growing research interest in deepfake detection and is suitable for academic publication and broader deployment in trust-critical environments.

	(%)	(%)	(%)	(%)	(%)
Model	Accurcy	Precision	Recall	F1-Score	AUC
SVM	79.4	78.5	76.2	77.3	0.81
Logistic regression	82.1	80.9	79..5	80.2	0.83
CNN	88.6	86.9	87.2	87.1	0.90
LSTM	90.3	89.5	88.6	89.0	0.96
CNN-LSTM (Proposed)	94.8	94.1	93.6	93.8	0.96

4. Discussion of Model

1.Comparison with Baseline Models

The proposed CNN-LSTM model significantly outperforms both traditional machine learning and standalone deep learning approaches in the task of fake audio detection. Traditional classifiers such as Support Vector Machines (SVM) and Logistic Regression achieved accuracy scores of 79.4% and 82.1%, respectively. While standalone CNN and LSTM architectures showed improved results—88.6% and 90.3%, respectively—the hybrid CNN-LSTM model reached a notable accuracy of 94.8%. This enhanced performance is primarily due to the synergy between the CNN's ability to learn discriminative spectral features from Mel-spectrograms and the LSTM's capability to model long-term temporal dependencies in the audio data. The model also demonstrated strong discrimination ability with an AUC of 0.96, confirming its effectiveness in real-world classification scenarios with imbalanced data.

2. Analysis of Precision, Recall, and F1-Score

Evaluation metrics such as Precision (94.1%), Recall (93.6%), and F1-Score (93.8%) highlight the model's reliability in differentiating fake from real audio. High precision means the model minimizes false positives, making it suitable for applications where flagging real audio as fake can have serious consequences—such as in forensics or authentication systems. Likewise, the high recall score indicates that the model can detect a majority of fake audio attempts, which is crucial in preventing misinformation or malicious audio manipulation. The balance between these two metrics, shown by the F1-Score, reflects the model's well-rounded performance across various deepfake scenarios.

3. Impact of CNN and LSTM Layers

The CNN layers play a key role in capturing spectral features that are often altered in deepfake audio, such as frequency artifacts introduced by synthesis methods. By processing Mel-spectrogram inputs, these layers identify local acoustic inconsistencies and patterns. On the other hand, the LSTM layers excel at tracking the temporal structure of speech signals, learning patterns across time. Since fake audio often exhibits inconsistencies in prosody, rhythm, or transitions, the LSTM component helps in identifying these flaws effectively. The combination ensures that both short-term and long-term characteristics of the audio are learned during training, improving detection accuracy.

4. Generalization and Robustness

The model's performance was evaluated on benchmark datasets, including the **SceneFake dataset from Kaggle**, to validate its generalization ability. Despite the diversity in audio manipulation techniques across these datasets, the CNN-LSTM model maintained high and consistent performance. This suggests that the model is robust to different styles of synthetic audio and capable of adapting to previously unseen types of fake content. Such robustness is essential for real-world deployments where the nature of fake audio continues to evolve.

5. Limitations and Future Directions

Despite its strong results, the CNN-LSTM model has certain limitations. The hybrid architecture demands moderate to high computational resources, making it less optimal for deployment in real-time or edge environments without optimization. Furthermore, since the model heavily depends on the quality of Mel-spectrogram representations, performance may degrade in the presence of significant background noise or distortion. Future enhancements could include **attention mechanisms**, **denoising pre-processing**, or **transformer-based architectures** to improve robustness and reduce latency in real-time scenarios.

5. Conclusion

This project introduced a deep learning-based CNN-LSTM architecture for the detection of deepfake audio using spectrogram-based features. Given the growing threat posed by fake audio content in media, authentication, and security domains, the need for reliable detection systems is critical. The CNN layers helped extract localized spectral features from audio spectrograms, while the LSTM layers effectively modeled the temporal structure of speech, identifying manipulations across time

Extensive evaluations on **SceneFake** dataset demonstrated the superiority of the proposed model compared to traditional and standalone methods. The model achieved high scores in accuracy, precision, recall, and AUC, confirming its effectiveness for real-world deployment. Its adaptability across different datasets also highlights its potential for general use in detecting emerging audio deepfakes.

While the current model shows excellent performance, further work can be directed toward reducing computation, enhancing noise robustness, and integrating real-time detection capabilities. This study lays a solid foundation for future research and applications aimed at ensuring trust and security in audio-based digital communication.

7.**References:**

1. T. Kinnunen, et al., “ASVspoof 2019: Automatic Speaker Verification Spoofing and Countermeasures Challenge,” *Proc. Interspeech 2019*, 2019.

2. K. Gong, et al., "FakeAVCeleb: A novel audio-visual dataset for deepfake detection," *IEEE Transactions on Information Forensics and Security*, 2021.

3. D. Snyder, et al., “Mel-Frequency Cepstral Coefficients for Speaker Recognition,” *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019.

4. Y. Kong, et al., "Spectrogram-Based Audio Classification," *IEEE Signal Processing Letters*, 2020.

5. H. Hassanpour, et al., “CNN for Audio Deepfake Detection: Convolutional Approaches for Manipulated Audio Identification,” *IEEE Access*, 2021.

6. N. Srivastava, et al., “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” *Journal of Machine Learning Research*, 2014.

7. M. Todisco, et al., “Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification,” *Computer Speech & Language*, 2017.

8. “Audio Deepfake Detection Using Deep Learning” IEEE R. Anagha1, A. Arya2, V. Hari Narayan3, S. Abhishek4 and T. Anjali5 2023

9. "Audio Deepfake Approaches" IEEE 2023 OUSAMA A. SHAABAN 1, REMZI YILDIRIM2, AND ABUBAKER A. ALGUTTAR

10. "Deepfake Audio Detection via MFCC Features Using Machine Learning" b2023 AMEER HAMZA1, ABDUL REHMAN JAVED 2,3, (Member, IEEE),FARKHUND IQBAL 4, (Member, IEEE), NATALIA KRYVINSKA 5,AHMAD S. ALMADHOR 6, (Member, IEEE), ZUNERA JALIL 2,AND ROUBA BORGHOL