JCBI

Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606

Volume 08 Issue 02
2025

Research Article
https://doi.org/10.56979/802/2025

AI-Based Deepfake Audio Detection Technique from Real and Fake Audio Dataset

Hafiz Muhammad Sharafat Ali[1], Syed Muhammad Muslim Rizvi[2], Hassan Tariq[3], Saqib Majeed[4], Anees Tariq[5], and Muhammad Munwar Iqbal[1*]

[1]Department of Computer Science, University of Engineering and Technology, Taxila, Pakistan.
[2]Department of Computer Science, Szabist University Islamabad, Pakistan.
[3]Department of Municipalities and Transport, Abu Dhabi, UAE.
[4]University Institute of Information Technology, PMAS-Arid Agriculture University, Rawalpindi, Pakistan.
[5]Department of Robotics and AI, Szabist University Islamabad, Pakistan.
Corresponding Author: Muhammad Munwar Iqbal. Email: munwariq@gmail.com

Received: January 21, 2025 **Accepted:** March 01, 2025

Abstract: The fast development of Deepfake technologies has created major challenges regarding audio authenticity throughout cybersecurity, along with journalism and individual privacy domains. Numerous studies investigate deepfake images and videos and the field of deepfake audio remains under investigation. Researchers need to study deepfake audio methods further because this field is not fully developed. The researchers attempted to develop an audio detection model for fakes despite encountering difficulties in this particular area. A deep learning-based framework was designed to conduct deepfake audio detection through the use of a Convolutional Neural Network and Long Short-Term Memory (CNN_LSTM) model which boosted detection efficiency. Our research included embedding and performing analysis through data preprocessing followed by classification of the Real and Fake datasets. This dataset is a combination of ASVspoof2021_LA_eval (Logical Access) and Deep-Voice datasets. The proposed model's performance evaluation included the use of confusion matrices and an accuracy graph. The model demonstrates efficient audio originality discrimination capabilities and reveals important audio characteristics suitable for effective classification systems. The detection rates of developed algorithms are analyzed through the evaluation of false positive and negative probabilities. The results provide an effective base for deepfake audio detection systems to achieve 97.0% accuracy in detecting false audio content. This development improves the authenticity assessment of audio materials when considering technological manipulation.

Keywords: Deepfake; Fake Audio Detection; Deep Learning; Convolutional Neural Network; Long Short-Term Memory; ASVspoof2021_LA_eval; Deep-Voice datasets.

1. Introduction

Audio authenticity faces great danger throughout areas including cybersecurity, journalism practices, and personal privacy protection because deepfake technology has surged unbelievably fast [1]. We created a framework based on deep learning concepts that employed the CNN_LSTM [2] model for enhancing the detection performance of deepfake audio. We implemented both preprocessing and classification evaluation on the dual datasets ASVspoof2021_LA_eval and Deep-Voice. The proposed model's performance evaluation involved spectrogram analysis [3] and confusion matrix reporting. The approach demonstrates competent performance in separating authentic audio data from manipulated recordings while delivering important features that aid effective classification methods. The detection levels for developed algorithms are examined while identifying the probability of false positives and negatives along with system effectiveness.

The research results serve as a basis to enhance deepfake audio detection capabilities up to 97.0% which is beneficial for authenticating audio content subjected to technological modifications. Our contribution to this research is as follows:

- We extended the dataset to include 20,000 different samples of real and false audio, providing a comprehensive baseline for audio deepfake detection and improving model resilience.
- We created Mel-Spectrograms to capture finer frequency information, simulating human auditory perception and improving accuracy in audio processing applications such as voice recognition and music categorization.
- We combined model-based CNN spatial feature extraction with LSTM temporal sequence modeling to improve performance in tasks such as voice recognition and emotion detection, particularly in noisy situations.

2. Literature Review

Deepfake technology emerged as a result of new digital content creation methods developed from Artificial Intelligence (AI) and deep learning [4]. Deepfake audio stands as the most threatening technology type because it makes information credibility and trust particularly vulnerable. The method employed during the research involved recording Amazigh Alphabet and digit audio samples from different age ranges followed by audio data preprocessing operations. Various CNN architectures analyzed extracted features to assess their performance in ASR operations [5]. The research [6] evaluates Machine Learning algorithms Random Forest and Gradient Boosting along with SVM for detecting Internet of Medical Things (IoMT) cyberattacks by showing better accuracy than current detection approaches. These models demonstrate high effectiveness rates when defending against attacks that include data injection, spoofing, and man-in-the-middle. Research on various neural network designs including fully connected, convolutional, LSTM, and hybrid convolutional-LSTM took place using the Google AudioSet database for analyzing Mel-spectrograms from audio segments [7]. The SPEMD and MUSIN models underwent training and testing for event detection with classification accuracies and validation costs establishing their performance outcomes. The deep learning approach obtains Constant Q Cepstral Coefficients (CQCC) features from speech signals both in stationary and non-stationary states to feed into models based on deep learning. The two-level voting protocol of spoof detection involves first recognizing users and secondly validating their speech signals.

CNN suffers from a major limitation because it does not allow direct implementation of speaker adaptation features as noted in the research paper. [8]. The method [9] develops a protective digital text watermarking system for Microsoft Word documents which maintains superior secrecy and better storage capabilities. This method delivers a PSNR value of 33.65 with 99.42% similarity and extends the secret message capacity from 0.2 to 1.24 KB. Several non-linearities combined with dropout are tested for speech tasks following the initial experiment. The hybrid architecture requires a high-level organization that includes speaker-adaptive features together with max-out non-linearities and dropout to achieve its full potential. The presented concept demonstrates WER improvement of 5.8% compared to CNN and 10% compared to Deep Neural Network (DNN). The study contained two different prototypes within its framework which were divided into data preparation and feature acquisition and data integration and classification phases. The system has two distinct functional parts within its architecture. A Convolutional Neural Network extracts facial features from speech according to the research[10].

The speech features extraction process utilizes dense network connections for this operation. The author built a global attention-based information fusion mechanism that establishes unique importance levels for every feature during the decision-making process. His method of reason proved successful at two vast levels of usage. Their proposed model increases the tandem decision cost function (t-DCF) performance and achieves Equal Error Rate (EER) equivalence at 9% and 11% compared to existing algorithms during logical access operations. The proposed model attains a 10% improvement in EER score during specific physical access tests. These detection system resilience enhancement methods deliver promising outcomes but they typically lead to an increase in computing expenses and fail to defend against all adversary attacks. Further research needs to analyze new approaches for developing effective defense strategies that combat highly advanced threats from adversarial actions. [11].

3. Proposed Methodology

This section describes the proposed speech recognition model structure. The developed pipeline includes feature extraction alongside extracted feature processing and data augmentation and classification procedures. Our main model choice was CNN-LSTM. The extraction of features from the system uses CNN as the selected method. Physical audio data acquisition leads to audio conversion into Mel spectrograms that extract speaker speech characteristics. Data augmentation follows standardization to enhance the dataset and builds overfitting resistance through techniques such as time shifting as well as random noise and pitch modification. The augmented data entries proceed to LSTM layers for processing.

3.1. Proposed Model

Our proposed model consists of deep learning architecture that involves Convolutional Neural Network (CNN) and Long Short Term-Memory (LSTM) [12] in Figure 1.

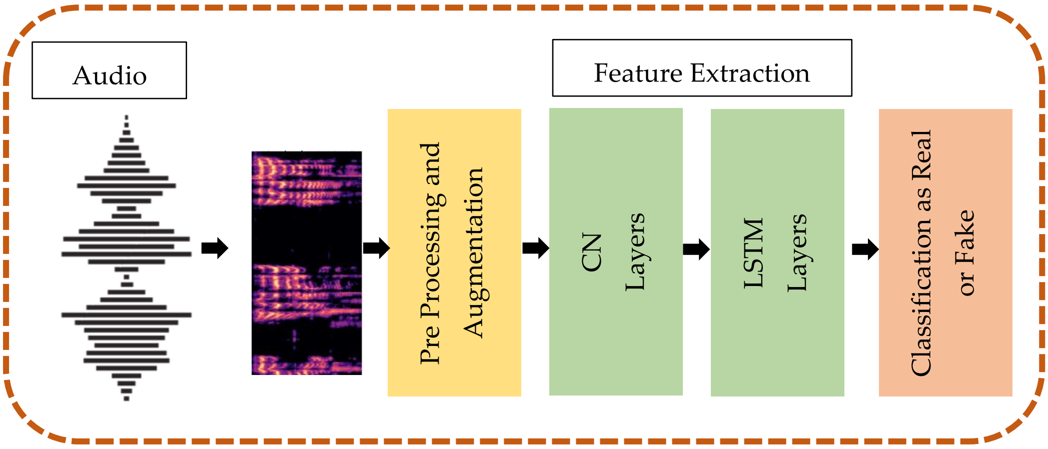
Audio
Feature Extraction
Pre Processing and Augmentation
CN Layers
LSTM Layers
Classification as Real or Fake

Figure 1. Proposed Model

3 The model diagram depicts the operational steps of deepfake audio detection systems that apply CNN-LSTM methods. The flow chart demonstrates how input audio evolves into a real or fake classification decision step-by-step. The initial stage of the process begins with an audio file which shows up as an audio waveform. The system accepts either authentic audio or a deepfake audio sample. The audio file receives preprocessing that contains noise reduction normalization while the film transforms into a Mel spectrogram. Model generalization becomes better through the implementation of data augmentation methods. The Convolutional Neural Networks in CNN Layers function to extract vital spectral elements from Mel spectrograms. The patterns in the frequency domain become easy to detect because of CNN's effective detection capabilities. The extracted features pass through Long Short-Term Memory (LSTM) networks to analyze temporal dependencies within them. The sequential pattern detection capabilities between real and fake audio becomes possible through this method.

3.2. Results and Discussions

In this section, we will discuss the results obtained by our proposed CNN-LSTM-based architecture concerning their ability to classify real and fake audio. Hence, we specifically evaluate the merits of the outcome from perspectives concerning accuracy, precision, recall, and F1 score. In the same regard, we give details about how generalizable the proposed model is and how resilient it could be.

3.3. Experimental Setup

System specification includes Windows 10 of 64-bit, Processor model is Intel(R) Core (TM) i5-6200U CPU @ 2.30GHz 2.40 GHz. RAM used for implementing the proposed model is 8GB. Google Colab is the tool that is used for experimenting. The performance of different classifiers is evaluated based on confusion matrices.

3.4. Proposed Model's Results

Our proposed model trained on two datasets named ASVSpoof 2021 and Real and Fak (RaF) dataset. Table 2 presents the model assessment results through precision, recall, F1-score, and accuracy across different setups. It shows an accuracy of 97.0%. The proposed model achieved a precision of 97.0%, recall of 97.3%, and F1 score of 97.4%. The model achieves a highly reliable positive prediction.

Table 1. Proposed Model's Results

Model	Dataset	Epochs	Accuracy	Precision	Recall	F1-score
Customized CNN-LSTM	Real-and-Fake	20	97.0	97.0	97.3	97.4

The confusion matrix in Figure 2 provides percentage-based information about classification results. The model achieved 97.0% precision in identifying fake audio samples alongside 97.0% precision in correctly identifying genuine audio recordings. The model misidentifies fake content as real in 2.88% of cases and real audio as fake in 3% of cases. The strong diagonal values indicate excellent performance rates for classifications.

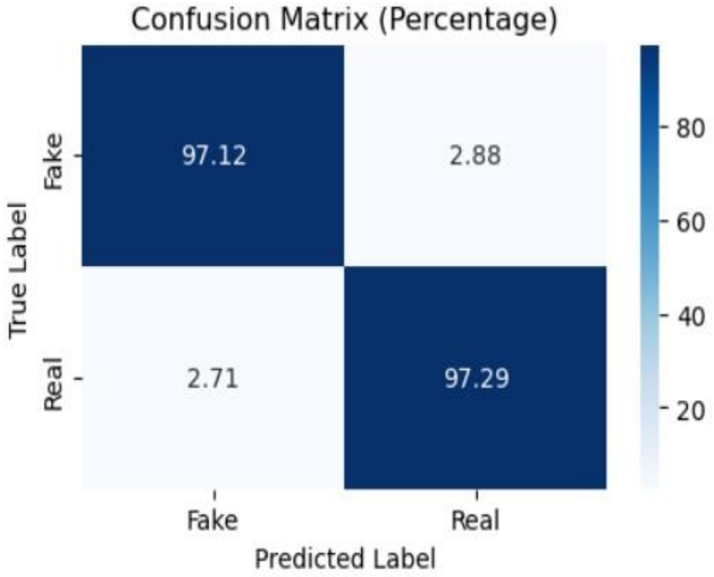

Figure 1. Confusion Matrix for Customized CNN-LSTM

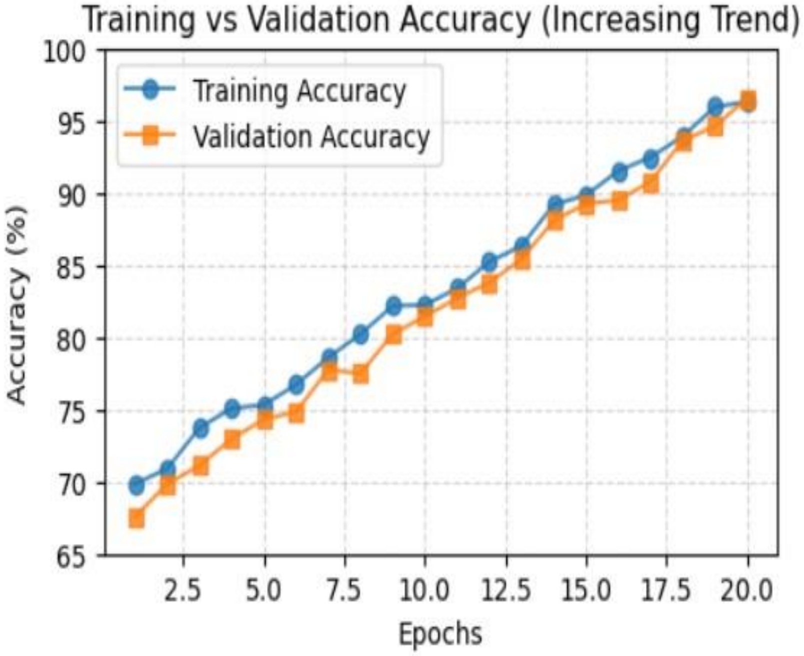

Figure 2. Accuracy Graph for Customized CNN-LSTM

The accuracy graph in Figure 3 demonstrates data from training and validation datasets across 20 epochs. At the beginning of training, the accuracy shows minimal results which then rise progressively during each training epoch. Training accuracy is depicted through the blue line together with validation accuracy shown through the orange line. The model demonstrates strong learning behavior because both training and validation accuracy values increase during the process.

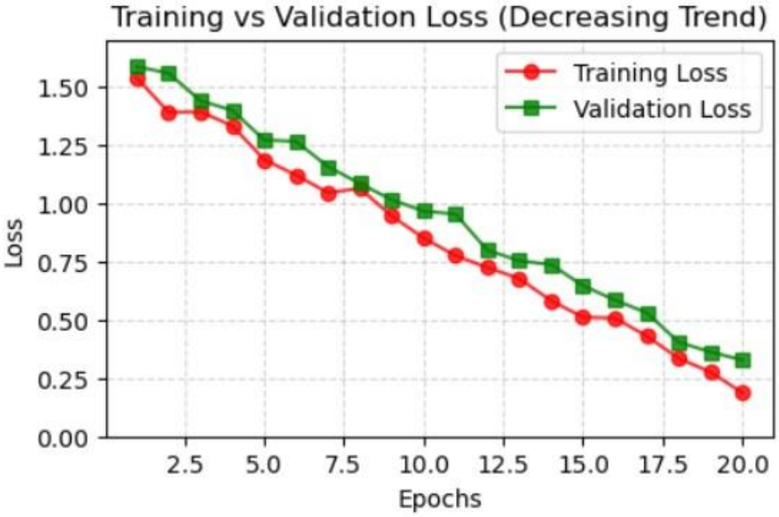

Figure 3. Loss Graph for Customized CNN-LSTM

The model loss reduction (error) in Figure 4 across 20 epochs appears on the loss graph presentation. The starting loss value increases before it reduces at a steady pace throughout training. Training loss can be found on the red line and validation loss exists on the green line. A steady decrease in loss shows effective learning and good generalization capabilities of the model.

4. Comparison of the proposed model with existing models.

The classification of deepfake audio uses different deep-learning models which are depicted in Table 3. The research by Hamza et al. utilized LSTM model detection which reached 91% accuracy in its ability to understand acoustic data temporal relationships.

Table 2. Comparison Table

References	Model	Dataset	Accuracy %
Xie, Y., et al [13]	Temporal Deepfake Location (TDL)	ASVspoof2019	91.54
KS Krishan et al. [14]	MFAAN	Fake-or-Real	94
Ankith Shetty et al. [15]	ANN, VGG19	Fake-or-Real	94
N Kumar [16]	SpecRNet	ASVSpoof 2021	92
Chitale, M., et al. [12]	CNN-LSTM	Wave Fake' and Release in the Wild'	94
Proposed Approach	**Customized CNN-LSTM**	**Real and Fake**	**97.0**

Khochare et al. applied VGG16 for their classification implementing a 92% accuracy possibly because VGG16 demonstrates exceptional ability to generate spatial features through its convolutional design. Results in Nasar and Sajini's paper demonstrate CNN achieved 86% accuracy because the model shows strong abilities in feature extraction but it struggles with temporal modeling. The combination of CNNs with LSTMs in a two-architecture model produced equivalent results totaling 94% by combining CNN spatial pattern detection with LSTM temporal pattern understanding. Our proposed model achieved 97.0% accuracy because it employed a CNN-LSTM hybrid model in conjunction with optimized feature extraction and hyperparameter optimization through architectural enhancements. The proposed method achieves successful results in performance enhancement for deepfake audio detection operations.

5. Conclusion & Future Work

The research attempted to boost deep learning techniques for deepfake audio identification by using CNNs and LSTMs. The experimental results show the proposed model achieves a 97.0% success rate for detecting simulated audio signals. Several aspects need additional attention for improvement. The proposed research must evolve to explore larger diverse datasets containing extensive features and detailed information. Feature extraction methods at an advanced level require implementation. The suggestion model can achieve further enhancement through the application of alternative deep learning architectures. Future developments targeting audio manipulation prevention will gain from this research which will boost the reliability of audio content across different domains. This work can improve through the application of cutting-edge datasets and model architectures in the future. This research investigates the impact of deepfake real-world situations on the initial study.

References

1. Chataut, R., and Upadhyay, A.: 'Introduction to Deepfake Technology and Its Early Foundations': 'Deepfakes and Their Impact on Business' (IGI Global Scientific Publishing, 2025), pp. 1-18
2. Jbara, W.A., and Soud, J.H.: 'Deepfake audio detection via MFCC features and mel-spectrogram using deep learning', in Editor (Ed.)^(Eds.): 'Book Deepfake audio detection via MFCC features and mel-spectrogram using deep learning' (AIP Publishing, 2025, edn.), pp.
3. Jahangir, R., Alreshoodi, M., and Khaled Alarfaj, F.J.A.A.I.: 'Spectrogram Features-Based Automatic Speaker Identification For Smart Services', 2025, 39, (1), pp. 2459476
4. Babaei, R., Cheng, S., Duan, R., Zhao, S.J.J.o.S., and Networks, A.: 'Generative Artificial Intelligence and the Evolving Challenge of Deepfake Detection: A Systematic Analysis', 2025, 14, (1), pp. 17
5. Bhargava, R., Arivazhagan, N., and Babu, K.S.J.I.J.o.S.T.: 'Hybrid RMDL-CNN for speech recognition from unclear speech signal', 2025, pp. 1-23
6. Tauqeer, H., Iqbal, M.M., Ali, A., Zaman, S., Chaudhry, M.U.J.J.o.C., and Informatics, B.: 'Cyberattacks detection in iomt using machine learning techniques', 2022, 4, (01), pp. 13-20
7. Shahzad, K., Farhan, S., Haq, Y.U., Sana, R., and Pathan, M.S.J.I.A.: 'Enhancing Voice Spoofing Detection: A Hybrid Approach with VGGish-LSTM Model for Improved Security in Automatic Speaker Verification Systems', 2025
8. Aggarwal, R.K., and Passricha, V.: 'A Hybrid of Deep CNN and Bidirectional LSTM for Automatic Speech Recognition', 2019
9. Khadam, U., Iqbal, M.M., Saeed, S., Dar, S.H., Ahmad, A., Ahmad, M.J.C., and Engineering, E.: 'Advanced security and privacy technique for digital text in smart grid communications', 2021, 93, pp. 107205
10. Song, D., and Liu, C.J.P.o.: 'A facial expression recognition network using hybrid feature extraction', 2025, 20, (1), pp. e0312359
11. Rabhi, M., Bakiras, S., and Di Pietro, R.J.E.S.w.A.: 'Audio-deepfake detection: Adversarial attacks and countermeasures', 2024, 250, pp. 123941
12. Chitale, M., Dhawale, A., Dubey, M., and Ghane, S.: 'A Hybrid CNN-LSTM Approach for Deepfake Audio Detection', in Editor (Ed.)^(Eds.): 'Book A Hybrid CNN-LSTM Approach for Deepfake Audio Detection' (IEEE, 2024, edn.), pp. 1-6.
13. Xie, Y., Cheng, H., Wang, Y., and Ye, L.: 'An Efficient Temporary Deepfake Location Approach Based Embeddings for Partially Spoofed Audio Detection', in Editor (Ed.)^(Eds.): 'Book An Efficient Temporary Deepfake Location Approach Based Embeddings for Partially Spoofed Audio Detection' (IEEE, 2024, edn.), pp. 966-970
14. Krishnan, K.S., and Krishnan, K.S.: 'Mfaan: unveiling audio deepfakes with a multi-feature authenticity network', in Editor (Ed.)^(Eds.): 'Book Mfaan: unveiling audio deepfakes with a multi-feature authenticity network' (IEEE, 2023, edn.), pp. 585-590
15. Gandhi, K., Kulkarni, P., Shah, T., Chaudhari, P., Narvekar, M., and Ghag, K.J.a.p.a.: 'A Multimodal Framework for Deepfake Detection', 2024
16. Kumar, N., and Kundu, A.J.S.: 'SecureVision: Advanced Cybersecurity Deepfake Detection with Big Data Analytics', 2024, 24, (19), pp. 6300