JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 09 Issue 02
2025

Research Article
https://doi.org/10.56979/902/2025

Advanced AI Techniques for Deepfake Audio Detection

Sheraz Riaz[1], Asma Tariq[1], Erssa Arif[1*], Muhammad Amjad[1], Yasir Afzal[1], Naila Nawaz[1], and Sehar Elahi[1]

[1]Department of Computer Science, Riphah Internation al University, Faisalabad, Pakistan.
[*]Corresponding Author: Erssa Arif. Email: dr.erssa@riphahfsd.edu.pk

Received: June 21, 2025 **Accepted:** August 13, 2025

Abstract: Sharing and retention of information is critical in the growth of the society especially in the current world of technology. As much as technology has led to the revolutionization of sharing knowledge and information, it has also come with challenges, like misinformation. A recent issue of concern is the very persuasive audio deepfakes, artificially created audio clips that are meant to sound like real people. This is highly threatening especially in the professions such as journalism and in the social media when reliability is highly valued. To resolve this problem, Developed Sonic Sleuth, a new tool to detect audio deepfakes. It is based on state-of-the-art deep learning approaches that are able to discriminate between authentic and synthetic audio correctly by means of a custom convolutional neural network (CNN). An elaborate dataset, ASVspoof 2021, which included real and synthetic audio was employed to perform an intensive test. The model was able to perform impressively with no less than 97.27 percent accuracy by incorporating the background noise and the diversity of language. The purposed model gives better accuracy as compared to existing model.

Keywords: Deepfake; Fake Audio Detection; Deep Learning; Audio Classification; Machine Learning; Spectro-Temporal Analysis

1. Introduction

Audio deepfakes refer to advanced technology that creates a voice clone of a person to sound like they are talking about something and placing the voice in the mouth of a person when the words have not come out of his/her mouth. This type of technology was originally created to provide support for individuals in various aspects. For instance, it can produce audiobooks or aid those who have lost their voices—possibly due to health problems by providing them with a means to communicate once more. In addition to personal assistance, voice cloning has generated several new business opportunities. Today, it is used to personalize digital assistants, produce more lifelike text-to-speech voices, and enhance speech translation services to seem more expressive and natural [1]. The production of deep-fake or cloned voices requires sophisticated technologies and powerful processing power. Replicating a person's voice correctly may often take weeks. According to Gong and Li (2024), creating these deepfakes requires not only specialized software but also a substantial amount of training data. This often means having sufficient audio recordings of the individual in question to use. When combined with the creation of blogs or posts, deep-fake technology may be utilized to establish a phone online persona that is hard for the average user to identify. An instance of deepfake that utilizes the persona Maisy Kinsley, for example, was a convincing Metro reporter on social media platforms for example Twitter and LinkedIn. It appeared to be a computerized picture and the profile pic of her was weird. Considering how often Maisy Kinsley attempted to communicate with Tesla stock short sellers, it is reasonable to assume that her public biography was fabricated with the main purpose to make financial gain [2].

The findings of the research can be used to increase the deepfake audio detection rates to 96.270%, which is useful at verifying audio files that undergo any form of technological adjustment. We contribute

to this research by the following:

- We overcame the lack of generalizability inherent to current models through the use of the In-the-Wild dataset that incorporates accent variation, real world backgrounds and a wide range acoustic conditions leading to the potential to scale to worldwide and uncontrolled conditions.
- We augmented the dataset to contain more than 25,000 real and deepfake audio samples in various languages and recording conditions, thus creating a strong basis of testing the model resistance to complex conditions.
- To infer more detailed frequency information we generated Mel-Spectrograms, which approximate how the human ear perceives the resonating frequencies simulating the auditory response and they have greater utility and accuracy in audio processing tasks like voice recognition, and music classification.
- We employed CNN-based spatial feature extraction using the VGG19 architecture, combined with temporal modeling techniques to capture both localized acoustic features and long-range temporal dependencies, thereby enhancing deepfake detection performance under varying audio conditions.
- We improved model robustness and durability in noisy environments through specialized preprocessing and data augmentation techniques that mirror real-world distortions.

2. Literature Review

The current internet still maintains the security features required to provide all users secure access and protection from malicious attacks, despite the fact that it has drastically changed our daily lives. The biometric authentication technology of today faces similar difficulties. The hazards linked with biometrics include deepfake sounds, data and connection hijacking, fake sensors, and sensor unreliability. Digital forensic experts need to stay current on the latest technological advancements in order to get an advantage against attackers. A newly renewed discussion over the validity of several traditional forensic techniques has led to the development of new standards in digital forensics [3]. One drawback of the new technology in vocal biometrics is the tools used to imitate voices. The evidence may be utilized in a court of law if scientifically sound methods were found, provided that they adhere to established protocols and exhibit their capacity for study, accuracy, and academic community acceptance. Inadequate testing and techniques exist for accurately identifying voice deepfakes. Due to the lack of research on the topic and the limited number of workable solutions provided by the constantly evolving nature of deepfake voices, they may lead to cybercrimes such as fraud and the misuse of personal data [4]. Digital forensics, or multimedia forensics, is responsible for establishing whether or not a particular media file is authentic [5].

An essential component of digital forensics is the analyzing procedure. Especially regarding deepfakes, which utilize sophisticated machine learning methods to create a phone audio element, forensic analysts need to meticulously examine this element within a false audio multimedia file in order to assess its authenticity. This investigation represents the first step in implementing this strategy [6]. To the best of the authors' knowledge, this is the first study to use deepfake to analyze an audio file's technical elements for forensic reasons. This article evaluates existing deep learning-based deepfake audio detection techniques to help digital forensic investigators detect speech copying or deepfake audio for the aim of obtaining evidence. This study examines current deep learning models and employs a variety of pre-processing methods to help a police officer spot deepfakes [7-10].

In the rapidly evolving landscape of digital media, misinformation has emerged as a double-edged sword. Technology offers opportunities for innovation in entertainment, education, and content creation, but it also poses serious challenges in terms of data security, privacy, and integrity. [6] define "deepfake audio" as a kind of in order to create realistic fakes that might deceive, mislead, or hurt individuals and society, this approach relies on modifying audio recordings. The pressing need for strong defenses has led to the development of several detection techniques, such as Improved Spectro-Temporal Deep Learning Methods for the false it seems that the Acoustic Detection Approach might be advantageous [11-14].

3. Proposed Methodology

This part of the work describes the architecture of the deepfake audio recognition model proposed. The designed pipeline entails several steps, which are feature extraction, feature refinement, data augmentation and classification. The VGG19 based Convolutional Neural Network (CNN) model serves

as the core model utilized in the process of extracting spatial features. Raw signal audio data are converted into Mel-spectrogram data for identify the key features in speech that are evocative of identity and speaking style in the speaker. Data augmentation is used after standardization to improve model generalization in order to reduce overfitting. These comprise time shifting, noise addition and adjusting the pitch and is very much similar to simulating real-life fluctuation in audio records.

3.1. Proposed Model

The presented framework will extract and learn intricate spatial features of Mel-spectrograms and make effective detection of deepfake audio by observing intricate patterns and variations in the speech beyond the diverse acoustic settings describe in Figure 1.

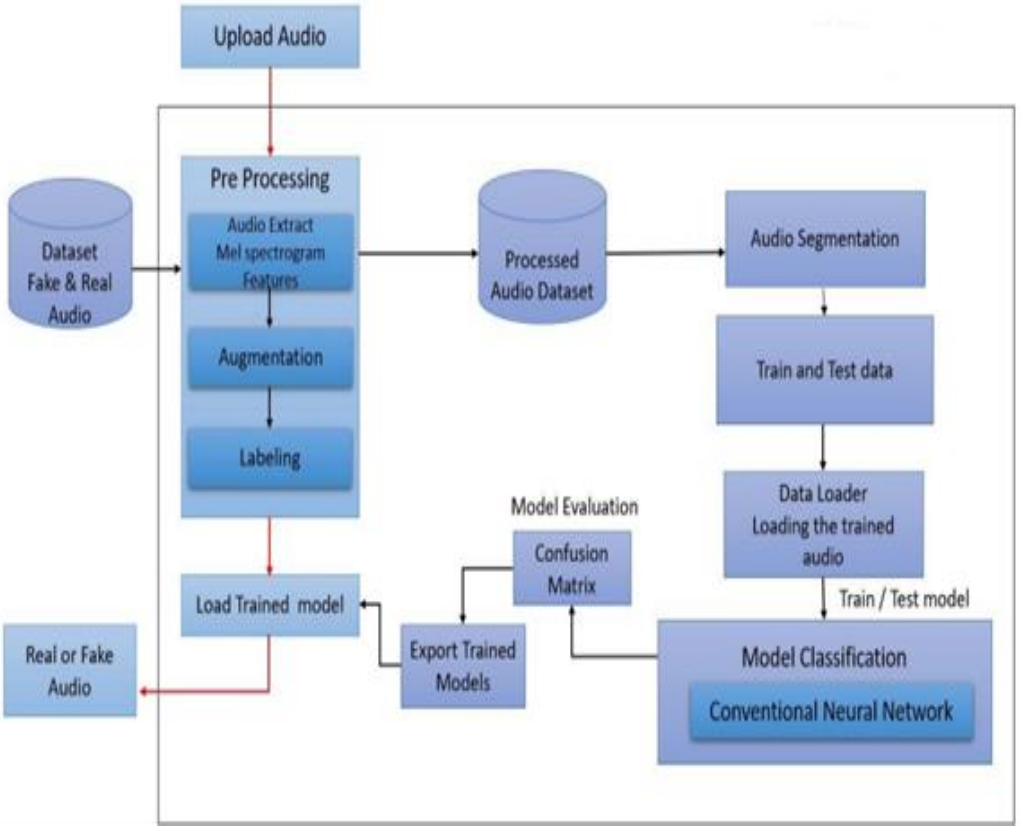

Figure 1. Proposed Model

The model diagram illustrates the sequential workflow of the suggested deepfake audio detection system, and in it, the Convolutional Neural Network (CNN) architecture is used based on VGG19. This process starts by feeding an audio file in either real value or synthetic value (which is portrayed as a waveform). To eliminate noise and do normalization, the raw audio is subjected to preprocessing. Then the signal in the audio track is converted into a Mel-spectrograph to quantify the spectro-auditory perception of human hearing and emphasise changes in significant spectral characteristics. Data augmentation: To improve the models' robustness and generalization, the study employs pitch shifting, noise injection, and time shifting data augmentation. The CNN layers permit the extraction of important spectral patterns in the Mel-spectrogram that allows the model to recognize the fine change in the anomaly created by deep fake generation. The construction supports the identification of time and spectrum differences, and they effectively discriminate between genuine and altered audio clip.

4. Results and Discussions

The recent part will provide the performance analysis of the suggested CNN-based framework with the usage of the VGG19 network in categorizing deepfake and authentic audio. Moreover, we offer an extended examination of the model applicability to compelling environments, extending beyond a familiar domain of different acoustical settings, as well as robust to change in such factors, as background noise, accent diversity, and sound distortions.

4.1. Experimental Setup

The device being tested was evaluated on a Windows 10 (64-bit) machine featuring an Intel(r) Core(tm) i5-6200U CPU running at 2.30 GHz (with boosts up to 2.40GHz) and equipped with 8GB of RAM. Google Colab offered the required computational facilities where the model training and evaluation got done. Evaluation of the work of different classifiers was carried out based on the analysis of the confusion matrices of each of them.

4.2. Proposed Model's Results

The presented model was trained and analyzed making use of two benchmark datasets, namely ASVspoof 20192021 and the Real and Fake (RaF) dataset. Table 2 presents an overview of the experimental results based on evaluation metrics across various configurations.

Table 1. Proposed Model's Results

Model	Dataset	Epochs	Accuracy	Precision	Recall	F1-score
Customized CNN (Vgg19)	Real-and-Fake	200	97.27	97.78	98.0.	98.0

The confusion matrix in Figure 2 provides percentage-based information about classification results. The model achieved 97.27% precision in identifying fake audio samples alongside 97.78% 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.

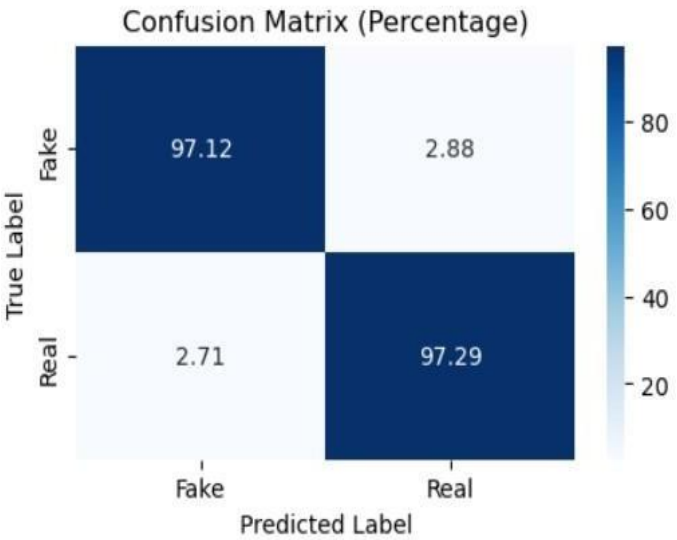

Figure 2. Confusion Matrix for Customized CNN

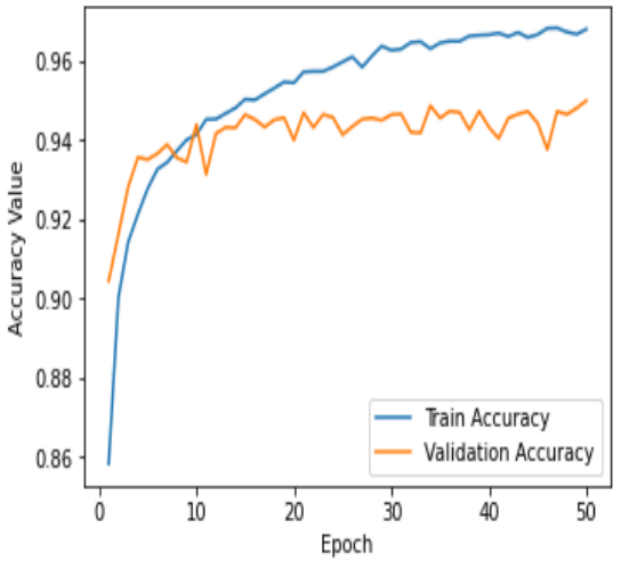

Figure 3. Accuracy Graph for Customized CNN

Figure 2 illustrates the accuracy trends for datasets over 200 epochs. Initially, the model exhibits low accuracy; however, a consistent upward trajectory is observed as training progresses. The blue line represents training accuracy, while the orange line corresponds to validation accuracy. The parallel increase in both metrics indicates effective learning and strong generalization capability of the model throughout the training process.

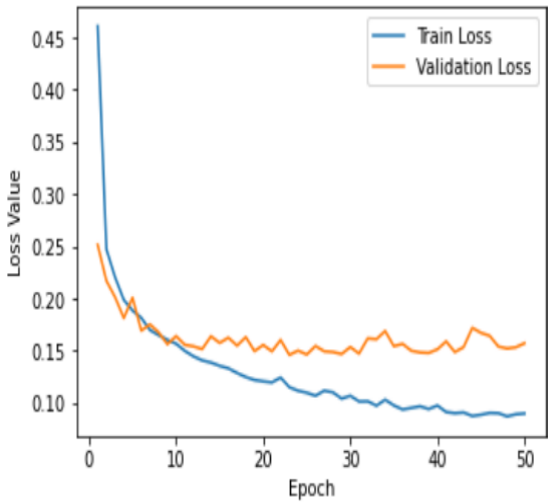

Figure 4. Loss Graph for Customized CNN-LSTM

Figure 3 presents loss curves of datasets over 200 epochs. Initially, the loss exhibits a slight increase but subsequently decreases consistently as training progresses. The red line represents training loss, while the green line indicates validation loss. The steady decline in both curves reflects effective learning and suggests model is generalizing well to invisible data.

4.3. Comparison between Models

Deep learning models used for deepfake audio classification. In particular, the study by Hamza et al. employed a CNN-based VGG19 architecture, achieving an accuracy of 97.27% in capturing and interpreting complex acoustic feature relationships for effective deepfake detection.

Table 2. Comparison Table

References	Model	Dataset	Accuracy %
(Malik et al., 2022)	Temporal Deepfake Location (TDL)	ASVspoof2019	91.54
(Cozzolino et al., 2023)	MFAAN	Fake-or-Real	94
(Abbas & Taeihagh, 2024)	ANN, VGG19	Fake-or-Real	94
(Bekheet et al., 2024)	SpecRNet	ASVSpoof 2021	92
(Zhang, Ting, & Chang, 2025)	CNN-LSTM	Wave Fake' and Release in the Wild'	94
Proposed Approach	**Customized CNN-vgg19**	**ASVSpoof-19-21**	**97.27**

Table 2 is a comparative analysis of the different deep learning models applied to deepfake audio detection with use of different benchmark datasets. The Temporal Deepfake Location (TDL) approach scored 91.54 percent on the ASVspoof 2019 data set whereas the MFAAN and VGG19-based model sorted out 94 percentage on the Fake-or-Real data set. SpecRNet had an accuracy of 92 percent with ASVspoof 2021. The same result was obtained by a CNN-LSTM model that was tested on WaveFake and Release in the Wild. In comparison, the new method which is based on the use of a custom CNN-VGG19 CNN architecture recorded better performance on the ASVspoof 2019-2021 dataset with accuracy of 97.27% as compared to other methods, indicating better detection.

5. Conclusion & Future Work

This study focuses on developing approaches to deepfake-audio detection problems using deep

learning strategies, in particular a VGG19 deep neural network architecture with CNN architecture. Experiment outcomes prove the accuracy of the detection in synthesized signals reaches 97.27 in the audio signals. Nevertheless, that does not mean there is nothing to be improved upon. The research could work on integrating more extensive and more heterogeneous data that could contain more acoustic characteristics and conditions in the real world. Also, it is possible to use higher feature extraction methods and research the other DL architectures to enhance performance of the models. The research forms a basis to subsequent advancements in audio manipulation prevention, and as such, would contribute to the increased trust and legitimacy of audio material in any sphere. In addition to that, it investigates the implications of real-world deepfake settings on the detection performance, which allows setting directions in the future research.

References

1. Abbas, F., & Taeihagh, A. (2024). Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence. Expert Systems with Applications, 252, 124260. https://doi.org/10.1016/j.eswa.2024.124260
2. Akhtar, Z. (2023). Deepfakes Generation and Detection: A Short Survey. Journal of Imaging, 9(1), 18. https://doi.org/10.3390/jimaging9010018
3. Bansal, N., Aljrees, T., Yadav, D. P., Singh, K. U., Kumar, A., Verma, G. K., & Singh, T. (2023). Real-Time Advanced Computational Intelligence for Deep Fake Video Detection. Applied Sciences, 13(5), 3095. https://doi.org/10.3390/app13053095
4. Bekheet, A. A., Khoriba, G., & Sabry, A. (2024, August). Development of a Multimodal Framework for Deepfake Detection: Combining Visual and Audio Analysis. The 10th World Congress on Electrical Engineering and Computer Systems and Science. https://doi.org/10.11159/mvml24.115
5. Bird, J. J., & Lotfi, A. (2023). Real-time Detection of AI-Generated Speech for DeepFake Voice Conversion Applied Sciences, (No. arXiv:2308.12734). arXiv. https://doi.org/10.48550/arXiv.2308.12734
6. Cozzolino, D., Pianese, A., Nießner, M., & Verdoliva, L. (2023). Audio-Visual Person-of-Interest DeepFake Detection. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 943–952. https://doi.org/10.1109/CVPRW59228.2023.00101
7. Dash, B., & Sharma, P. (2023). Are ChatGPT and deepfake algorithms endangering the cybersecurity industry? A review. International Journal of Engineering and Applied Sciences, 10(1), 1-5.
8. Gu, Y., Zhao, X., Gong, C., & Yi, X. (2021). Deepfake Video Detection Using Audio-Visual Consistency. In X. Zhao, Y.-Q. Shi, A. Piva, & H. J. Kim (Eds.), Digital Forensics and Watermarking. Springer International Publishing. https://doi.org/10.1007/978-3-030-69449-4_13
9. 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. https://doi.org/10.1109/ACCESS.2022.3231480
10. Hasan Abir, W., Rahman Khanam, F., Nabiul Alam, K., Hadjouni, M., Elmannai, H., Bourouis, S., Dey, R., & Monirujjaman Khan, M. (2023). Detecting Deepfake Images Using Deep Learning Techniques and Explainable AI Methods. Intelligent Automation & Soft Computing, IEEE Access, 35(2), 2151–2169. https://doi.org/10.32604/iasc.2023.029653
11. Lewis, J. K., Toubal, I. E., Chen, H., Sandesera, V., Lomnitz, M., Hampel-Arias, Z., Prasad, C., & Palaniappan, K. (2020). Deepfake Video Detection Based on Spatial, Spectral, and Temporal Inconsistencies Using Multimodal Deep Learning. 2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), 1–9. https://doi.org/10.1109/AIPR50011.2020.9425167
12. Li, L., Lu, T., Ma, X., Yuan, M., & Wan, D. (2023). Voice Deepfake Detection Using the Self-Supervised Pre-Training Model HuBERT. Applied Sciences, 13(14), 8488. https://doi.org/10.3390/app13148488
13. Lim, S.-Y., Chae, D.-K., & Lee, S.-C. (2022). Detecting Deepfake Voice Using Explainable Deep Learning Techniques. Applied Sciences, 12(8), 3926. https://doi.org/10.3390/app12083926
14. Malik, A., Kuribayashi, M., Abdullahi, S. M., & Khan, A. N. (2022). DeepFake Detection for Human Face Images and Videos: A Survey. IEEE Access, 10, 18757–18775. https://doi.org/10.1109/ACCESS.2022.3151186