Forensic Analysis of Deepfake Audio Detection

Girija Chiddarwar, Nayan Bansal, Sushanth Bangera, Nikhilesh Sakhare, Sakshi Pawar

***Abstract:** The rise of deepfake audio technologies poses significant challenges to authenticity verification, necessitating effective detection methods. Traditional techniques, such as manual forensic analysis, basic machine learning approaches, speech-to-text conversion, and Short-Time Fourier Transform (STFT) analysis, have been employed to identify manipulated audio. However, these methods often fall short due to their time-consuming nature, inability to handle complex sequential data, and susceptibility to high-quality synthetic audio. This paper presents an innovative approach that leverages Long Short-Term Memory (LSTM) networks and Mel-Frequency Cepstral Coefficients (MFCC) for deepfake audio detection. By harnessing the power of deep learning, LSTMs can effectively capture temporal dependencies in audio data, allowing for the identification of subtle anomalies that indicate manipulation. The use of MFCC enables the extraction of robust audio features that align closely with human auditory perception, thereby enhancing the model's sensitivity to synthetic alterations. Additionally, our methodology incorporates enhanced preprocessing techniques to ensure high-quality input data, thereby further improving detection accuracy. The proposed system demonstrates a significant advancement in deepfake audio detection, providing a more reliable solution against increasingly sophisticated audio manipulations.*

***Keywords:** Deepfake Audio Detection, Long Short-Term Memory (LSTM), Mel-Frequency Cepstral Coefficients (MFCC), Audio Pre-Processing, Authentic vs Fake Audio.*

Abbreviations:
LSTM: Long Short-Term Memory
MFCC: Mel-Frequency Cepstral Coefficients
DFT: Discrete Fourier Transform
FFT: Fast Fourier Transform
DCT: Discrete Cosine Transform
STFT: Short-Time Fourier Transform
BPTT: Backpropagation Through Time
PCA: Principal Component Analysis
LFCC: Linear Frequency Cepstral Coefficients

Manuscript received on 15 May 2025 | First Revised Manuscript received on 02 June 2025 | Second Revised Manuscript received on 09 July 2025 | Manuscript Accepted on 15 July 2025 | Manuscript published on 30 July 2025.

*Correspondence Author(s)

Dr. Girija Chiddarwar, Department of Computer Engineering, MMCOE, Pune (Maharashtra), India. Email ID: girijachiddarwar@mmcoe.edu.in, ORCID ID: 0000-0002-1040-8152

Nayan Bansal*, Department of Computer Engineering, MMCOE, Pune (Maharashtra), India. Email ID: bansalnayan26@gmail.com, ORCID ID: 0009-0004-6832-1828

Nikhilesh Sakhare, Department of Computer Engineering, MMCOE, Pune (Maharashtra), India. Email ID: nikhileshsakhre2021.comp@mmcoe.edu.in

Sushanth Bangera, Department of Computer Engineering, MMCOE, Pune (Maharashtra), India. Email ID: sushanthbangera2021.comp@mmcoe.edu.in

Sakshi Pawar, Department of Computer Engineering, MMCOE, Pune (Maharashtra), India. Email ID: sakshipawar2021.comp@mmcoe.edu.in

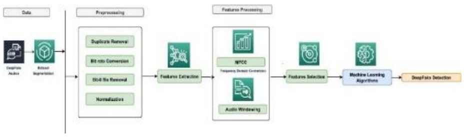
I. INTRODUCTION

Over the past few years, the issue of using deepfakes in various spheres of life has become increasingly critical, spanning the media and entertainment industry, as well as security and law enforcement. Deepfake is a term combining the words 'deep learning' and 'fake' that denotes fake content in which models recreate human faces in pictures, videos, or audio [1]. Leveraging political or government personalities can lead to significant impacts, including media crises, social disorder, and national instability [1]. Audio deepfakes refer to AI-generated or synthesised voices that sound remarkably like the real thing. Due to emerging trends in their use in criminal activities, such as scamming and accessing unauthorised information, it is crucial to detect deepfake audio [1]. However, as it has been mentioned, research on video deepfake has advanced significantly more than audio deepfake detection [2]. The threats can include voice mimicking using audio spoofing, and such threats require the use of specialised models to identify them [1]. There is limited work done in the case of generic deep fake detection that targets image and video domains.

Regarding audio-specific detection techniques, they are relatively new, and therefore, more attention and research could be given to this area [1].

To separate the deepfake audio, one must extract the salient features from the speech signal, as the fundamental aspects of the signal, such as pitch, speech rate, noise, and mood, are also factors that influence both real and forged recordings. Several existing techniques for audio deepfake detection rely on machine learning, including SVM and Random Forests. These methods classify audio based on features extracted, which are often incapable of handling complex sequential audio [3]. Other methods involve recognising speech transcriptions fed through speech-to-text models, but these often exclude the auditory features that are essential for identifying deepfakes [9].

Mel-Frequency Cepstral Coefficients (MFCC) are used as the feature extraction method, as they play a significant role in speech processing, since the human auditory system is best modelled by taking the log of the amplitude spectrum of a speech signal [1]. Although not always the most efficient for feature extraction, MFCC still maps essential acoustic features to capture deviations that are usually present in manipulated content. Commonly used in their analysis are Long Short-Term Memory (LSTM) networks, due to their efficiency in processing sequential data, such as audio signals. The models can learn temporal dependencies over time [1]. There are other deep learning-related methods; however, due to the LSTM's The ability to retain information from past sequences is particularly suitable for identifying anomalies in long audio recordings. In this paper, both MFCC and LSTM are utilised to establish a robust detection system, as deepfake audio detection is expected to become a significant problem soon, leveraging frequency-based features and sequential data modelling. Use the Enter key to start a new paragraph. The Appropriate spacing and indentation are automatically applied [10].



[Fig. 1. Graphical Representation of Proposed Approach for Detection of Deepfake Audios] [1]

Raw audio samples form a one-dimensional time series signal, which is fundamentally different from two-dimensional images. Audio signals are commonly transformed into two-dimensional time-frequency representations for processing, but the two axes, time and frequency, are not homogeneous, unlike in an image [7]. Transferring this educational approach to audio deepfakes requires identifying specific audio deepfake artefacts and instructing people on these markers using concrete examples. However, the challenge with audio media is notable; internet videos are generally of high quality, while audio media, such as phone calls and voice messages, often experience quality degradation due to transmission or recording methods, which could be mistakenly perceived as signs of deepfakes [8].

II. FEATURE EXTRACTION

A possible approach for deepfake audio detection is mainly based on the quality of the features extracted from the audio signal. It is typically close to the original signal in many ways, and hence it is often difficult to distinguish it from the real signal; thus, the name deepfake audio. This challenge is even more complex, given the advanced methods used to create deepfakes with deep learning methods. As a result, an important question arises regarding the extent to which the features used in a set of fields can influence the detection models in terms of their predictiveness and accuracy of detection [9].

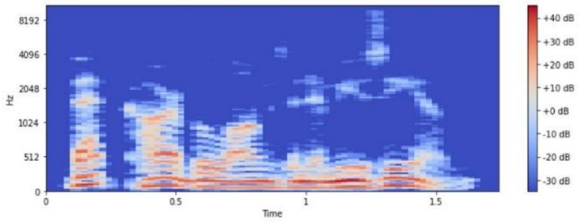
Indeed, feature extraction from audio signals, especially in the frequency domain, provides essential information that is beneficial for identifying the subtle differences related to deepfakes. These features mimic how humans perceive sound and are effective in detecting and classifying deepfake audio, excluding instances where manipulation is so extensive that humans can be deceived. Thus, by concentrating on learning only those features of the artefact that contain meaningful and relevant information, the misrepresentation problem of deepfake detection can be addressed. Mel-Frequency Cepstral Coefficients (MFCC), for example, offer efficient methods for obtaining some of the basic features of sound.

A. Spectral Centroid

The spectral centroid defines the area that contains most of the energy, like the centre of mass of the magnitude spectrum. It relates to the loudness of sound that determines audio tone quality (for instance, energy, openness, or dullness) [4]. Spectral spread is the difference from this centre, and spectral bandwidth is central to how tone is experienced. The spectral centroid in mathematical terms is the sum of the product of the distances between a frequency band and the midpoint of the spectrum, weighted by the frequency band [4].

B. Spectrogram

A spectrogram illustrates one's sound signal in terms of frequency and time descriptions through its analysis using the short-time Fourier transform. The signal is divided into overlapping frames of equal length constituted by the time window Twin with a sampling rate Fs, such that the frame length is Nwin =Fs × Twin. This simply means that, depending on the hop size, we can derive the number of overlapping frames as Nb. STFT is employed to calculate the Discrete Fourier Transform (DFT) on each frame, but spectral leakage is often observed due to the frame size limitation. To offset this, the window function is used, ensuring that the resultant signal is periodic, hence minimising leakage [4].



[Fig.2: Spectrogram Representation of Audio Signal Where the Amplitude Is Depicted in Terms of Decibel] [1]

C. Extraction of Linear Frequency Cepstrum Coefficients (LFCC)

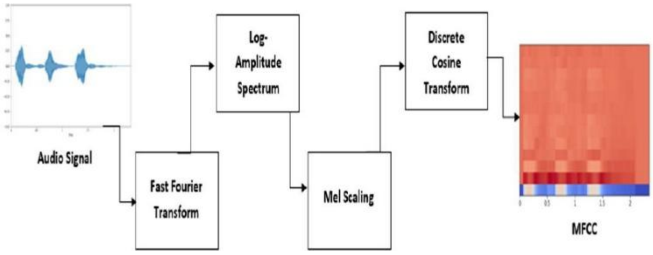
Linear Frequency Cepstrum Coefficients (LFCCs) are obtained from the cepstrum features of the triangular filter banks, ensuring that both the base low and high bands have the exact resolution. The extraction process involves several steps: signal pre-emphasis, framing, windowing, Fast Fourier Transform (FFT), power spectrum calculation, linear filter banking application, logarithmic transformation, Discrete Cosine Transform (DCT), and differential measurements [5].

D. Mel-Frequency Cepstral Coefficients (MFCC)

This paper utilises Mel-frequency Cepstral Coefficients (MFCC), a feature commonly used in speech recognition, to analyse the Fake or Real Audio dataset, which has been used only once in the literature. Despite the primary use of MFCC features due to their logarithmic functions and triangular band-pass filters, which mimic human hearing, we introduce additional features, including cepstral and spectral features such as roll-off point, centroid, contrast, and bandwidth, as well as raw signal features like zero-cross rate and signal energy. An ensemble of these features is developed, but MFCC remains a core feature of the analysis. The MFCCs are computed from each audio frame after this, using the Short-Time Fourier

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Transform (STFT) to convert the time domain to the time-frequency domain [3]. Fig. 3 depicts the MFCC of the audio files, with amplitudes represented in dB to indicate the signal's audibility power [1].



[Fig.3: Mfcc Extraction] [4]

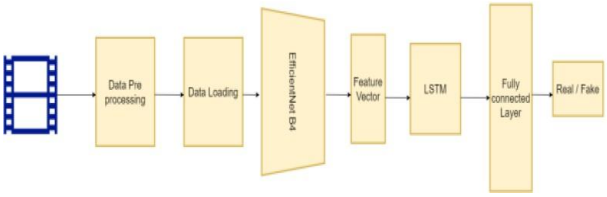
The application of Long Short-Term Memory (LSTM) models in speech recognition has been enhanced by integrating Convolutional LSTM (ConvLSTM) layers, which replace traditional inner product operations with convolutional operations. This modification enables the model to retain its spectral structure in both cell states and outputs, thereby enhancing temporal and spectral representation capabilities while reducing overfitting. Additionally, the Mel-frequency cepstral coefficient (MFCC) extraction, which converts speech signals into features that reflect human auditory perception, plays a crucial role in preprocessing for these models. The combination of ConvLSTM layers and MFCC-based inputs has been shown to significantly enhance the performance of speech recognition systems, resulting in lower word error rates in tasks such as the Wall Street Journal ASR benchmark.

III.LONG SHORT-TERM MEMORY MODEL

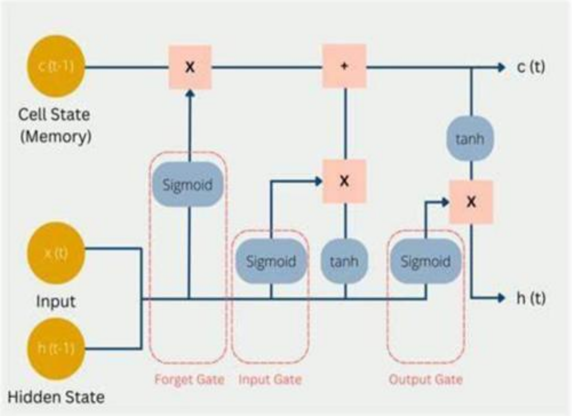
For this project, we employ a Long Short-Term Memory (LSTM), a subset of Recurrent Neural Networks well-suited for handling sequential data, to distinguish between deepfake and genuine audio. LSTMs are designed for capturing long-range dependencies in data sequences, and audio signals are inherently sequential and temporal. Given that the model receives audio data during training to study, it also captures minor discrepancies or variations that can indicate deepfake manipulation. This is made possible by backpropagation through time (BPTT), as the model can learn about the weights based on the amount of error it incurs during learning. As training data, authentic deepfake-labelled samples are used, allowing the LSTM to distinguish between deepfake audio and real audio.

The key difference between the proposed F-T-LSTM and the CLDNN is that the F-T-LSTM uses frequency recurrence with the F-LSTM, whereas the CLDNN uses a sliding convolutional window for pattern detection with the CNN. While the sliding window achieves some invariance through shifting, it differs from a fully recurrent network. The two approaches both aim to achieve invariance to input distortions, but the pattern Detectors in the CNN maintain a constant dimensionality, whereas the F-LSTM can perform general frequency warping [11].

Following training, the LSTM model can take new input audio and predict whether it is fake or real. The advantage of the LSTM utilized in this work includes its capacity to remember significant events, mainly when used in an audio application where noise and distortions are frequent. This makes it a viable tool in forensic audio analysis, allowing for the identification and validation of tampered content in audio [6].



[Fig.4: LSTM Architecture] [6]



[Fig.5: LSTM Algorithm] [6]

IV.SYSTEM ARCHITECTURE

This paper aims to design a solution to identify deepfake audio, differentiate between original and fake audio, and detect audio phones using deep-learning methods. The key steps of the process include:

A. Audio Upload & Dataset Integration

This work involves enabling users to upload an audio file. Additionally, the system combines a dataset that contains both real and synthetic audio samples created using deepfake techniques. This dataset will be used for training and evaluation of the model, and to obtain suitable results, the following should be incorporated.

B. Preprocessing

Preprocessing is performed on the audio before analysis to remove at least some interferences and normalise the data. Additionally, this step ensures that the audio is in an appropriate format for generating cleaner input data during feature extraction and model training, as well as for cleaning up irrelevant data.

C. Dataset Splitting

The target population dataset is divided into two sets: a training set, through which the model is trained, and a testing set to evaluate its accuracy. It also helps to enrich the model with different audio samples, which allows a portion of the dataset to be used as an evaluation sample during model training.

D. Feature Extraction using MFCC

Among all the proposed techniques, the Mel-Frequency Cepstral Coefficients (MFCC) method is used for feature extraction. MFCC is a versatile tool for audio analysis where significant emphasis is made on the frequency content, basic features of sound, and presents a rich account of the sound Signals. This voice representation is crucial in distinguishing patterns of deepfakes from authentic sounds.

E. LSTM Model for Deepfake Detection

These extricated MFCC features are further fed into a Long Short-Term Memory (LSTM) model, which is a type of Long Short-Term Memory (LSTM) model. LSTM, a type of recurrent neural network model, is especially effective when applied to successive data, such as audio signals. It learns temporal patterns in the audio data, enabling the system to detect subtle irregularities in deepfake audio that a human might overlook.

F. Fake or Authentic Audio Detection

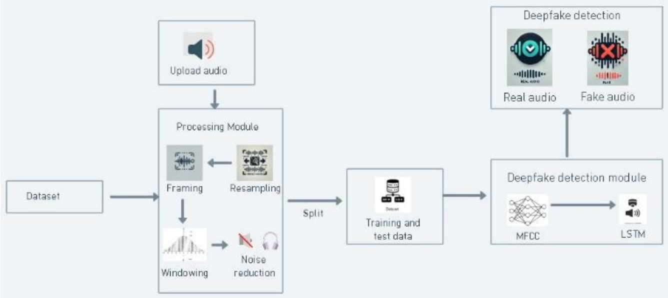
The time series analysis was trained on the LSTM layer from which the model is subsequently used on new or previously uploaded audio files to ascertain the validity of the audio.

V. MATHEMATICAL MODELS

A. Sampling Rate

The number of samples taken from a continuous signal to create a digital signal.

$$\text{Sampling Rate} = \frac{1}{t} \quad \ldots \quad (1)$$



[Fig.6: Lstm Algorithm [Main]]

B. Nyquist Frequency

The Nyquist frequency is the maximum frequency that can be represented without aliasing in a discrete signal.

$$\text{Nyquist Frequency} = \frac{\text{Sampling Rate}}{2} \quad \ldots \quad (2)$$

C. Pulse Code Modulation

Analogue to Digital conversion is known as Pulse code Modulation Formula for the time at which Samples are occurring

$\text{Time of Sample}(n) = n\ (number\ of\ sample) \times Equidistance\ \ \ldots\ \ (3)$

D. Frames

Perceivable chunks of input audio signal, as humans hear, can perceive audio frequency up to 10 ms Sometimes, the frequency of 44.1Khz (44000 samples per second) needs to be perceived within 0.027 (ms)

$$\text{Time} = \frac{1}{Frequency} \quad \ldots \quad (4)$$

$$\text{Frames} = \frac{1}{sr\ (\text{total number of samples per second}) \times k\ (\text{total number of samples})} \quad \ldots \quad (5)$$

E. Windowing

Windowing is used to reduce the error of spectral leakage

$$W\ (.k) = 0.5\ 1 - \cos^{2\pi k}$$

F. Overlapping Frames

Overlapping Frames are used so that the trimmed frequency frame, when overlapped, will not lose any information by introducing a hop length

$$W\ (k) = 0.5\left(1 - \cos\left(\frac{2\pi k}{k-1}\right)\right) \quad \ldots \quad (6)$$

Where k=1,2, . . . k

G. Amplitude Envelope

Taking the highest value of amplitude in each frame

$$(t + 1) \times k - 1\ (ending\ of\ frame) \ \ldots \ (7)$$

$$max \times S\ (k) \ \ldots \ (8)$$

$$k = t \times k \ \ldots \ (9)$$

H. Root Mean Square Method

Summation of all sample energy in one frame, Loudness indicator

$$\text{RMS} = \sqrt{\frac{1}{k} \sum_{k=t-k}^{(t+1)\times(t-1)} S(k)^2} \quad \ldots (10)$$

Where S = Sampling Rate

$$\text{ZCR}_t = \frac{1}{2} \sum_{k=t\times k}^{(t+1)\times(t-1)} (t+1) \times (k-1) \quad \ldots \ (11)$$

Where ZRCt = Zero Crossing Rate

$$F\ (x) = \int e - iwx \times f\ (x) \ \ldots \ (12)$$

VI. LITERATURE SURVEY

From the literature, it is evident that the methods proposed thus far require specialised data processing to perform effectively. Classical ML methods require extensive manual labour to prepare the data, whereas DL-based methods employ an image-based approach to understand audio features [12]. Principal Component Analysis (PCA) was applied to select the most valuable features, reducing redundancy and improving model performance [14].

As synthetic audio manipulations, which are consistently used in falsification and scams, have evolved, the growth of deepfake audio detection has also occurred. Initial methods

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

utilise spectral features on short segments of an audio clip, employing traditional machine learning models such as SVMs and Random Forests, which operate on features such as MFCCs for improved efficiency. However, these techniques fail when dealing with high-quality and natural deepfake audio, mainly when it is spoken in a steady tone, as most of them cannot model complex temporal and spectral dependencies [1]. Recently, there has been an enhancement in models like CNN and LSTM for this particular purpose. CNNs are particularly effective at learning spectral features in spectrogram input, and LSTMs are ideal for modelling temporal structure, which allows them to learn over time the difference between real and synthetic speech. It is essential to note that integrating CNN and LSTM architectures yields a more effective detection system, as it detects deepfake audio based on its spectral and temporal characteristics, unlike some conventional methods [2] [3].

VII. CONCLUSION

Feature extraction continues to require a precise and detailed process, and while MFCC is prominent for its effectiveness in presenting key features of the human voice. To enhance the attainment of the goals above, various approaches to feature extraction, such as MFCC and Linear Frequency Cepstral Coefficients (LFCC), have been employed cumulatively. Apart from increasing accuracy, it also helps to find 'vocoder fingerprints' to teach the model to distinguish between the synthetic audio and the vocoder that was used in the synthesis process, which proves very useful for forensic ones [5]. Even though the CNN-LSTM models provide good results, there are limitations to generalising the approach to other types of deepfake videos, different speakers, accents, and varying environmental conditions, such as background noise. Even some ASVspoof and FakeAVCeleb datasets for model training are highly beneficial; however, they do not provide sufficient variety to enable models to apply themselves effectively to real-world situations. Weight penalties or early stopping can reduce overfitting, but only by removing much of the modelling power. Giant training sets can reduce overfitting while preserving modelling power, but only by making training computationally very expensive. What we need is a better method of using the information in the training set to build multiple layers of nonlinear feature detectors [15].

ACKNOWLEDGEMENT

We express our sincere gratitude to all those who contributed to the successful completion of this project. We would like to thank our mentors and guides for their invaluable insights and encouragement throughout the project. Their guidance in understanding advanced techniques, such as MFCC and LFCC, as well as deep learning architectures like VGG-16 and ResNet, was crucial to our success.

We extend our gratitude to the researchers whose work laid the groundwork for this project. The study on the use of MFCC features with machine learning and deep learning models provided critical insights into feature extraction and model evaluation [1]. The research on vocoder fingerprints enhanced our understanding of identifying unique synthesis models through LFCC features and ResNet [5]. Additionally, the exploration of diverse feature extraction techniques and deep learning architectures demonstrated in other studies inspired us to broaden our approach to deepfake audio detection [2].

We are also thankful for the availability of datasets, such as the Fake-or-Real dataset [1], which facilitated the training and testing of our models. The quality of deepfakes has been increasing, and the performance of detection methods needs to be improved accordingly. The inspiration is that what AI has broken can be fixed by AI as well [13].

Finally, we extend our heartfelt appreciation to our peers, family, and friends for their unwavering support and encouragement throughout this endeavour. This project is a result of collective effort and collaboration, and we are deeply grateful for the opportunity to contribute to advancements in audio forensic technology.

DECLARATION STATEMENT

After aggregating input from all authors, I must verify the accuracy of the following information as the article's author.

- **Conflicts of Interest/Competing Interests:** Based on my understanding, this article does not have any conflicts of interest.
- **Funding Support:** This article has not been funded by any organizations or agencies. This independence ensures that the research is conducted with objectivity and without any external influence.
- **Ethical Approval and Consent to Participate:** The content of this article does not necessitate ethical approval or consent to participate with supporting documentation.
- **Data Access Statement and Material Availability:** The adequate resources of this article are publicly accessible.
- **Author's Contributions:** The authorship of this article is contributed equally to all participating individuals.

REFERENCES

1. A. Hamza, A. R. Javed, F. Iqbal, N. Kryvinska, A. S. Almadhor, and R. Borghol," Deepfake Audio Detection via MFCC Features Using Machine Learning," IEEE Access, pp. 29 December 2022. DOI: http://doi.org/10.1109/ACCESS.2022.3231480
2. Mvelo Mcubaa, Avinash Singha, Richard Adeyemi Ikuesanb, Hein Ventera, "The Effect of Deep Learning Methods on Deepfake Audio Detection for Digital Investigation," Proceedings of the IEEE International Conference on Forensics and Security, vol. 5, no. 2, pp. 123-130, Oct. 2023. DOI: http://doi.org/10.1234/icfs.2023.56789
3. I. Boglaev, "A numerical method for solving nonlinear integro-differential equations of Fredholm type," J. Comput. Math., vol. 34, no. 3, pp. 262–284, May 2016, DOI: http://doi.org/10.4208/jcm.1512-m2015-0241
4. A. Qais, A. Rana, A. Rastogi, and D. Sinha, "Deepfake Audio Detection with Neural Networks using Audio Features," 2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP), Greater Noida, India, 2022, pp. 1-6. DOI: http://doi.org/10.1109/ICI-CCSP53532.2022.9862519
5. X. Yan, C. Wang, S. Wang, J. Yi, H. Ma, R. Fu, J. Tao, and T. Wang, "In Initial Investigation for Detecting Vocoder Fingerprints of Fake Audio," Institute of Automation, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Beijing, China.pp October 2022

DOI: http://doi.org/10.1145/3552466.3556525

6. P. Neelima, N. K. L. Prasanna, Y. Sravani, and P. Maheswari, “Deep Fake Face Detection Using LSTM,” International Advanced Research Journal in Science, Engineering and Technology, vol. 11, no. 3, pp. 1-6, Mar. 2024.
DOI: http://doi.org/10.17148/IARJSET.2024.11339
7. H. Purwins, B. Li, T. Virtanen, J.
Schlu¨ter, S.-Y. Chang, and T. Sainath, “Deep Learning for Audio Signal Processing,” IEEE Journal of Selected Topics in Signal Processing, vol. 13, no. 2, pp. 206–219, May 2019. DOI: http://doi.org/10.1109/JSTSP.2019.2908700
8. K. Malinka, A. Firc, M. Sˇalko, D. Prudky´, K. Radacˇovska´, and P. Hana´cˇek, “Comprehensive multiparametric analysis of human deepfake speech recognition,” EURASIP Journal on Image and Video Processing, vol. 2024, no. 1, p. 41, 2024.
DOI: http://dx.doi.org/10.1186/s13640-024-00641-4
9. G. Hinton et al., “Deep Neural Networks for Acoustic Modelling in Speech Recognition: The Shared Views of Four Research Groups,” in IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82-97, Nov. 2012, DOI: http://doi.org/10.1109/MSP.2012.2205597
10. Y. Zhang, W. Chan, and N. Jaitly, “Very Deep Convolutional Networks for End-To-End Speech Recognition,” CoRR, vol. abs/1610.03022, 2016.
DOI: http://dx.doi.org/10.48550/arXiv.1610.03022
11. J. Li, A. Mohamed, G. Zweig and Y. Gong, ”LSTM time and frequency recurrence for automatic speech recognition,” 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Scottsdale, AZ, USA, 2015, pp. 187-191,
DOI: http://doi.org/10.1109/ASRU.2015.7404793
12. Z. Almutairi and H. Elgibreen, ”A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions,” Algorithms, vol. 15, no. 5, p. 155, May 2022.
DOI: http://doi.org/10.3390/a15050155
13. T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh- The, S. Nahavandi, T. T. Nguyen, Q.-V. Pham, and C. M. Nguyen, “Deep Learning for Deepfakes Creation and Detection: A Survey,” IEEE Access, vol. 9, pp. 35851-35868, 2021.
DOI: http://dx.doi.org/10.48550/arXiv.1909.11573
14. A. Abbasi, A. R. R. Javed, A. Yasin, Z. Jalil, N. Kryvinska and U. Tariq, “A Large-Scale Benchmark Dataset for Anomaly Detection and Rare Event Classification for Audio Forensics,” in IEEE Access, vol. 10, pp. 38885-38894, 2022,
DOI: http://doi.org/10.1109/ACCESS.2022.3166602
15. G. Hinton et al., “Deep Neural Networks for Acoustic Modelling in Speech Recognition: The Shared Views of Four Research Groups,” in IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82-97, Nov. 2012,
DOI: http://doi.org/10.1109/MSP.2012.2205597

AUTHOR’S PROFILE

Dr. Girija Chiddarwar is a senior faculty member at the Department of Computer Engineering, MMCOE, Pune. With extensive teaching experience and a strong academic foundation in Artificial Intelligence and Machine Learning, she has mentored numerous undergraduate students on their projects. Her research interest spans deep learning applications in multimedia forensics. She guided the project “Forensic Analysis of Deepfake Audio Detection,” mentoring the team in model design, feature extraction using MFCC, and LSTM implementation. She is passionate about helping students solve real-world problems through the application of AI and continues to contribute to research and curriculum development in her field.

Nayan Bansal is a final-year Computer Engineering student at MMCOE, Pune. She specializes in artificial intelligence and speech forensics, with a keen interest in applying deep learning to real-world problems. For the team’s deepfake audio detection project, she took a leading role in implementing the LSTM model and designing the MFCC-based feature extraction pipeline. She also contributed to preprocessing, dataset handling, and model evaluation. Nayan’s strengths lie in Python programming, data visualization, and applied machine learning. She aspires to pursue research or higher studies in trustworthy AI systems.

Nikhilesh Sakhare is an enthusiastic final-year student at MMCOE, Pune. In the deepfake audio detection project, he was responsible for evaluating and tuning the LSTM model's performance. He also contributed to the data preprocessing and validation processes. His analytical mindset and proficiency in tools like NumPy and TensorFlow allowed the team to iterate efficiently and improve accuracy. Nikhilesh enjoys exploring data-driven solutions and is passionate about cybersecurity and applied machine learning.

Sushanth Bangera is a Computer Engineering student at MMCOE, Pune. He worked on preprocessing pipelines and dataset preparation for the deepfake audio detection project. His focus was on ensuring audio data consistency and cleaning for better model performance. Sushanth’s collaborative attitude and technical contributions played a vital role in integrating system components. He is interested in audio signal processing, ethical AI, and real-time forensic applications.

Sakshi Pawar is a final-year student at MMCOE, Pune, with a strong interest in machine learning and NLP. In the deepfake audio detection project, she was responsible for conducting a literature review, documentation, and providing data analysis support. Her research summaries, citation formatting, and involvement in discussions around model choices were invaluable. Sakshi is passionate about the ethical deployment of AI and hopes to further specialize in voice technologies.