A Survey on Deep Fake Audio Detection Using Deep Learning

Sai Retish Reddy, Prakash Kumar Sarangi, Shravan Nikhil, M A Jabbar and Sukanya Mekala
Department of CSE (AI&ML), Vardhaman College of Engineering, Hyderabad, Telangana, India

Keywords: Deep Fake Audio Detection, Audio Spoofing, wav2vec, Mel-Frequency Cepstral Coefficients (MFCC), Voice Cloning, Deep Learning (DL).

Abstract: The rise of synthetic voice generation techniques such as deepfakes and voice cloning, has raised concerns about potential misuse in areas like fraud, misinformation, and identity theft. Deep Learning has emerged as an effective solution for detecting deep fake audio. Features from audio files were extracted and analyzed using Mel-Frequency Cepstral Coefficients (MFCC) to illustrate the differences between real and fake audio. By examining patterns in the audio data, models can learn to distinguish subtle discrepancies. Results indicated that a Custom Architecture outperformed, with the VGG-16 architecture yielding the best results for MFCC image features. Using Deep Learning approaches, this paper presents an overview of deep fake audio detection methods, highlighting the importance of this study and the potential uses of Deep Learning Models in the fight against deep fake audio.

1 INTRODUCTION

Deepfake technology involves the creation of synthetic media, including audio, video, images, and text, that are typically designed to appear genuine Shaaban et al., (2023).

The evolution of artificial intelligence, particularly in deep learning, has made the generation of realistic synthetic audio, often known as "deepfake audio," increasingly prevalent. This technology allows for the replication of speech that closely resembles an individual's voice, which raises significant concerns over its use in identity theft, misinformation, and fraud. Consequently, distinguishing between genuine and synthetic audio has become critical, particularly in sensitive domains such as cybersecurity, journalism, and content moderation Rana et al., (2022).

Deepfakes pose threats to privacy, security, and authenticity. While much research has concentrated on deepfake video detection, achieving higher accuracy, audio spoofing remains an urgent issue that requires specialized detection models Hamza et al., (2022).

Audio deepfake detection presents unique challenges compared to image deepfakes, as visual cues like unnatural facial movements are absent. Synthetic audio convincingly mimics the subtleties of human speech, such as tone, rhythm, and accent, complicating detection efforts Hamza et al., (2022). Moreover, rapid advancements in audio synthesis technology underscore the need for equally sophisticated detection strategies. In order to improve the accuracy of the classification models, many approaches use machine learning algorithms such as Random Forest, Decision Tree and SVM. Our experiments utilized the Fake-or-Real Dataset, which included four sub-tests for thorough evaluation.

Deep learning offers a compelling approach to this issue by enabling automatic learning and feature extraction from extensive audio datasets. Techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have demonstrated efficacy in areas such as speech recognition and natural language processing and can be adapted for the detection of synthetic audio. Training models on real and fake audio data allows neural networks to capture nuanced patterns and inconsistencies that may be challenging for human listeners to detect. Figure 1 illustrates the deep fake classification. Figure 2 shows the general deep learning models.

Reddy, S. R., Sarangi, P. K., Nikhil, S., Jabbar, M. A. and Mekala, S.
A Survey on Deep Fake Audio Detection Using Deep Learning.
DOI: 10.5220/0013911900004919

In *Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies (ICRDICCT'25 2025) - Volume 4*, pages 283-289
ISBN: 978-989-758-777-1

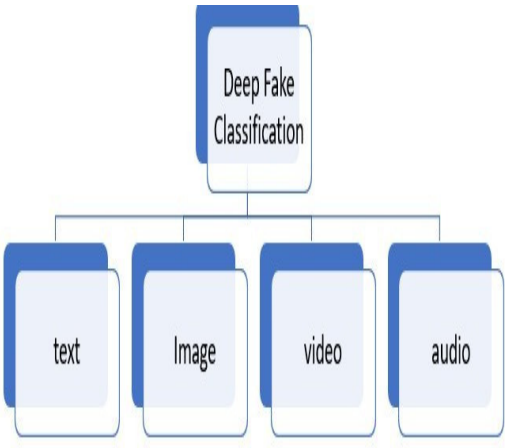

Figure 1: Deep Fake Classification.

A key component in deepfake audio detection is the use of Mel Frequency Cepstral Coefficients (MFCCs), which represent the short-term power spectrum of sound. These coefficients, along with various audio features, enable models to discern acoustic differences between authentic and synthetic audio. CNNs are particularly proficient in detecting frequency domain variations, while RNNs excel in recognizing temporal patterns, which are essential for spotting irregularities in speech rhythm or pronunciation. The field of deep learning-based fake audio detection is rapidly advancing and holds significant promise for addressing the growing threats posed by synthetic audio technology.

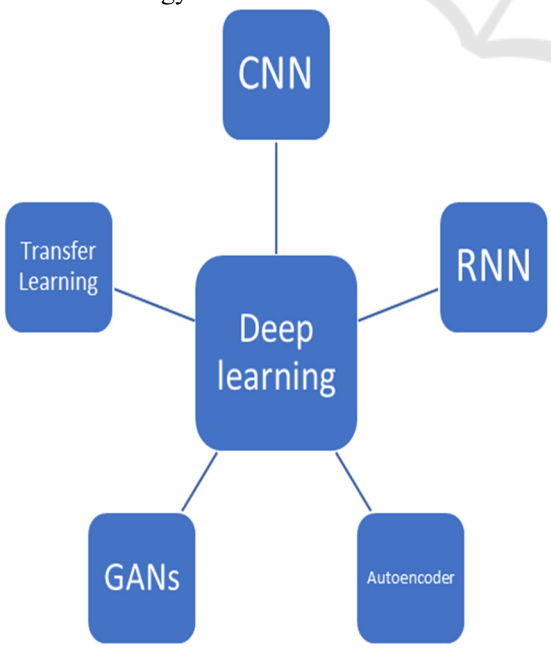

Figure 2: General Deep Learning Models.

Deep learning techniques are vital for the detection of deep fake audio. Various approaches can be utilized to achieve this. One prominent method is the use of Convolutional Neural Networks (CNNs), which excel in analyzing audio spectrograms. CNNs can extract features from these visual representations that signal the presence of manipulated audio. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, are another useful method. These networks are good at identifying the temporal correlations in audio data. LSTMs can identify unique patterns that are often present in deep fake audio. Autoencoders also play a role in detecting deep fake audio. These neural networks are designed to compress and then reconstruct audio data. When trained on authentic audio samples, autoencoders can recognize deviations that suggest the audio has been altered. Moreover, Generative Adversarial Networks (GANs) can be utilized in this context. A GAN consists of a generator, which creates new audio samples, and a discriminator, which assesses how realistic these samples are. By exposing the discriminator to real audio data, it learns to distinguish between genuine and deep fake audio. Lastly, Transfer Learning can enhance deep fake audio detection. This technique involves starting with a pre-trained neural network, which has already been learned from a vast audio dataset, and adapting it for the specific task of identifying deep fake audio.

1.1 Need for Deep Fake Audio Detection in Digital Security

The emergence of deep learning methods has brought about considerable progress in creating highly realistic synthetic media, particularly in the realm of audio deep fakes. While these innovations hold promising potential for various creative fields, they also pose significant risks, especially in areas where the authenticity of audio is vital. Deep fake audio can be misused for harmful purposes, including fabricating voice recordings in legal matters, generating fraudulent voice commands for security violations, or impersonating people in telephone scams.

Voice-based authentication systems, often utilized in banking and other sensitive sectors, are particularly susceptible to attacks using deep fake audio. These threats not only jeopardize individual privacy but also pose risks to institutional and national security. Additionally, the alteration of public figures' voices in political discussions can

erode trust in media and contribute to widespread misinformation.

In light of these increasing dangers, it is critical to develop effective systems for detecting deep fake audio. Such detection mechanisms are vital for preserving the authenticity of voice communications, preventing identity theft, and ensuring the reliability of digital audio content. Integrating deep fake detection technologies within digital security frameworks can help reduce these threats and safeguard against the advancing challenges of audio-based deception.

2 DEEP FAKE DETECTION USING DEEP LEARNING

2.1 Deepfakes

The evolution of deepfake technology has extended to audio, particularly concerning virtual assistants and other forms of computer-generated sound that are increasingly prevalent in our everyday lives Shaaban et al., (2023). The use of artificially created or altered audio content presents a major risk to society, as it can create trust issues when people struggle to differentiate between genuine and fake material.

Deep fake audio, being a form of synthetic media, has developed extremely fast with advances in advanced machine learning models like Generative Adversarial Networks (GANs) and voice cloning technologies. These advances facilitate the production of extremely realistic-sounding audio recordings that can be almost impossible to distinguish from genuine voices. Though these technologies hold promising uses in entertainment, accessibility, and other creative applications, they also pose real risks of fraud, disinformation, and security breaches. Deep fake audio misuse for voice phishing, impersonation fraud, and creating evidence for legal or political purposes is of particular concern.

As the threat from audio deception grows, it has become essential to develop effective detection systems. Deep fake audio detection models are now vital for maintaining the reliability of voice-based systems, including biometric authentication, legal documentation, and media content. By identifying fraudulent audio, these detection systems play a crucial role in ensuring the credibility of digital materials and mitigating the risks associated with manipulated voice data. This paper examines different deep learning methods used to detect deep fake audio and proposes a methodology utilizing the Wav2Vec2 model.

2.2 Applications of Deep Fake Audio Detection

Deep fake audio detection has applications across multiple domains where voice authenticity is critical. In the financial sector, many banks and institutions use voice biometrics for authentication in their systems Rana, M.S. and Sung, A.H., (2020) A deep fake audio attack could be used to impersonate a person, gaining unauthorized access to sensitive information or funds. Similarly, in cybersecurity, the rise of voice-controlled devices has made it imperative to detect audio impersonation to prevent unauthorized control over these systems.

Beyond security, the detection of deep fake audio is crucial in the media industry Patel et al., (2023). The use of fake voices to manipulate the public's perception, especially involving public figures, can lead to misinformation and destabilization of trust in media sources. Political campaigns, journalism, and entertainment all rely on the authenticity of audio content, and the development of effective detection models is vital in maintaining credibility and preventing the spread of false narratives.

2.3 Deep Learning Technique for Deep Fake Audio Detection

Several deep learning techniques are used to detect deep fake audio. Convolutional Neural Networks (CNNs), though originally developed to analyze images, have been applied to scan spectrograms of audio signals. Focusing on the visual patterns within these spectrograms, CNNs can detect abnormalities common to deep fake audio, like unsmooth frequency patterns. Also, Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are often utilized to handle sequential audio data and encode the temporal relations that get manipulated in creating fake audio.

More recently, Transformers and speech recognition models like Wav2Vec2 have emerged as powerful tools for deep fake detection. These models are capable of processing raw audio data and extracting deep representations that can highlight subtle anomalies present in synthetic audio. Autoencoders are also employed in

detection systems, as they can learn the patterns of genuine audio and flag deviations, indicating a potential deep fake. Together, these deep learning techniques form the backbone of modern deep fake audio detection systems.

2.4 Deep Learning Approaches

Convolutional Neural Network: Convolutional Neural Networks (CNNs) are a powerful deep learning framework widely used for detecting deep fake audio. Like other neural network architectures, CNNs comprise an input layer, an output layer, and several hidden layers. In the realm of deep fake audio detection, these hidden layers play a crucial role in processing audio inputs and performing convolutional operations on these signals.

By conducting these operations, CNNs can identify significant features within audio that may indicate modifications or artificial origins. Instead of relying only on basic matrix multiplication, CNNs utilize non-linear activation functions, such as Rectified Linear Units (ReLU), which introduce non-linearity into the model. This enhancement allows them to capture complex patterns in audio data. Furthermore, CNNs incorporate pooling layers to reduce the dimensionality of the data while preserving essential information. Techniques like average pooling can be used to summarize feature maps, thus lowering computational demands and improving the model's generalization ability. Through these architectural elements, CNNs can effectively detect subtle artifacts and inconsistencies in audio signals, making them a strong option for identifying deep fake audio.

Recurrent Neural Network: The Recurrent Neural Network (RNN) is an application of artificial neural networks that learns patterns from sequential data. It consists of multiple hidden layers, each characterized by its unique weights and biases. In an RNN, the nodes are connected in a directed cycle graph that sequentially processes data. This structure provides a recurrent hidden state, which effectively captures dependencies over time, enabling the network to manage temporal sequences Mary, A. and Edison, A., (2023).

Here, RNNs examine the temporal dependencies and patterns in audio waveforms so that they can detect speech nuances that could be indicative of manipulation. Through training on a dataset of both real and deep fake audio samples, RNNs are able to learn to detect slight inconsistencies in pitch, tone, and rhythm that are typical of synthetic audio. This ability allows RNNs to recognize real voices and artificially synthesized ones, thereby proving to be an effective resource against deep fake technology.

Transformers: The attention mechanism acts as a means of resource prioritization, allowing the model to concentrate on the most significant areas of an audio signal while suppressing the effect of irrelevant information. In deep fake audio detection, this approach helps the model emphasize important audio features, including speech patterns and changes in pitch, which are important for identifying synthesized alterations. Through focusing on these important characteristics, the model becomes more proficient at distinguishing between original and tampered audio and hence increases classification accuracy.

The Transformer architecture has proven effective in various fields. In computer vision, self-attention has become increasingly important as it is thought to generate superior deep feature representations through the calculation of weighted sums of features. Remarkable results have been achieved in numerous challenging computer vision tasks using self-attention techniques, alongside ongoing efforts to elucidate the workings of the self-attention mechanism.

Motivated by the successes in vision-related tasks, we propose that self-attention may also enhance efforts for detecting spoofed audio. In fact, adopting a Transformer- based framework has led to better performance outcomes. Some research has investigated the implementation of self-attention as a key mechanism to improve detection for partially fake audio.

Transformers leverage self-attention techniques to analyze audio data simultaneously, effectively capturing both local and global relationships within audio sequences. Unlike conventional recurrent neural networks (RNNs) that handle data sequentially, Transformers can examine complete audio segments at once, which makes them more efficient and effective in detecting complex patterns. Models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) can be repurposed for audio applications, enabling them to understand intricate relationships in speech and sound. When fine-tuned for deepfake audio detection, Transformers significantly bolster the model's ability to identify subtle inconsistencies indicative of synthetic audio, thereby enhancing overall detection performance.

3 METHODS FOR DETECTING DIFFERENT TYPES OF FAKE AUDIO

Detecting fake audio, particularly deep fakes, is a growing area of research and development. Various methods have been proposed to identify synthetic audio, leveraging advancements in machine learning, signal processing, and audio analysis. Below are some of the prominent methods used in this field

3.1 Detection of Synthesized Speech

Synthesized speech, generated using models like GANs or text-to-speech systems, presents one of the most common forms of deep fake audio. Detecting this type of audio often involves analyzing the unnatural tonal variations or background artifacts present in the synthesized audio Wani, T.M. and Amerini, I., (2023) Deep learning models such as CNNs applied to the spectrogram of the audio can reveal these discrepancies. Additionally, machine learning classifiers trained on synthesized versus real audio samples can distinguish between real and fake with high accuracy by recognizing the specific characteristics of speech generation algorithms.

In machine learning, a logistic regression model was designed to detect counterfeit audio files. To obtain the dataset, the authors used an imitation approach that entailed the extraction of entropy features from authentic and artificial audio samples. The model that was trained using this dataset achieved a remarkable accuracy of 98%. It is, however, noteworthy that preprocessing of the data has to be done manually to obtain the relevant features used by the model Rabhi et al., (2024).

3.2 Detection of Voice Cloning

Voice cloning, where a speaker's voice is imitated with high accuracy, poses a more challenging task for detection. Advanced neural network models, such as Wav2Vec2 or other transformer-based systems, can capture subtle details in the speaker's voice, making it possible to detect slight deviations from the original speaker's voice print Almutairi, Z.M. and Elgibreen, H (2023). These models analyze the prosody, pitch, and timbre of the audio to identify voice cloning attempts. Features like speaker embeddings, extracted using speaker recognition models, also play a crucial role in detecting cloned voices by comparing the speaker's identity with known voice profiles.

The Arabic-AD model on the Ar-DAD dataset achieved an impressive 97% accuracy Almutairi, Z.M. and Elgibreen, H (2023), while Deep Ensemble Learning (DEL) with Multi-LSTM on DF-TIMIT and Celeb-DF datasets yielded 89.73% Rana et al., (2022). The combination of wav2vec 2.0 models with light-DARTS on ADD 2022 Track 1 and ASVspoof 2019 LA datasets resulted in 80.4%, reflecting dataset complexity challenges Wang et al., (2022). A CNN integrated with a self-attention mechanism and ResNet achieved 89.4% accuracy on the ASVspoof 2021 dataset, emphasizing enhanced feature extraction Huang, L. and Pun, C.M (2024). Deep4SNet performed exceptionally on the H Voices dataset with 98.5% accuracy Rabhi et al., (2024), and a VGG-16 and LSTM-based model on the FOR-REREC dataset achieved 93% Hamza et al., (2022). Similarly, a CQT and MFCC-based model on FakeAVCeleb reached 94.15%, underscoring the value of frequency-based features the FG-LCNN model combined with ResNet achieved 95.93% accuracy on GitHub audio demos Mcuba et al., (2023), while ResNet18 on the ASVspoof 2021 DF dataset showed moderate performance with 74.21% Yang et al., (2024). Finally, a Siamese CNN with a Gated Recurrent Unit (GRU) on the ASVspoof 2019 dataset demonstrated 55% accuracy for the Siamese CNN model and a remarkable 99.8% for the Transformer- based model, highlighting the effectiveness of Transformers in audio deepfake detection Shaaban et al., (2023). These studies collectively reveal the progress and challenges in developing robust models for audio deepfake detection. Table 1 shows the deep learning models used for deep fake audio detection.

Table 1: Deep Learning Models used for deep fake audio detection.

S. No	Reference Number	DL Model Used	Dataset	Accuracy
1	Almutairi, Z.M. and Elgibreen, H., (2023)	Arabic-AD	Ar-DAD dataset	97%
2	Rana, M.S., Nobi., (2022)	Deep Ensemble Learning (DEL), Multi-LSTM	DF-TIMIT, Celeb-DF	89.73%
3	Wang, C., Yi, (2022)	wav2vec 2.0- base + light-DARTS,	ADD 2022 Track 1, ASVspoof 2019 LA	80.4%
4	Huang, L. (2024)	CNN, Self- attention mechanism, ResNet	ASVspoof 2021 dataset	89.4%
5	Rabhi, M., (2024)	Deep4SNet	H Voices	98.5%
6	Hamza, A., (2022)	VGG-16, LSTM	FOR-REREC DATASET	93%
7	Alshehri, A., Almalki, D., (2024)	CQT, MFCC based model	FakeAVCeleb	94.15%
8	Mcuba, M., Singh, (2019)	FG-LCNN, ResNet	audio demos. GitHub	95.93%
9	Yang, Y., (2024)	ResNet18	ASVspoof 2021 DF	74.21%
10	Shaaban, O.A., (2023)	Siamese CNN, Gated Recurrent Unit (GRU)	ASVspoof 2019	Siamese CNN: 55%, Transformer Model: 99.8%

4 CONCLUSIONS

This study investigated the application of deep learning methods for the detection of deepfake audio, with a particular emphasis on convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer models. Our findings reveal that deep learning approaches can effectively differentiate between genuine and altered audio, with transformers demonstrating the greatest potential. These models may serve as important resources in the fight against the increasing threat of deepfake audio. However, further research is essential to enhance their robustness and effectiveness in practical situations. Future studies could consider the incorporation of multimodal data, such as video in conjunction with audio, to develop more comprehensive deepfake detection systems. Additionally, addressing ethical considerations and advancing policy development will be crucial in tackling the challenges associated with deepfake technology.

REFERENCES

Shaaban, O.A., Yildirim, R. and Alguttar, A.A., 2023. Audio Deepfake Approaches. *IEEE Access*, *11*, pp.132652-132682.

Rana, M.S., Nobi, M.N., Murali, B. and Sung, A.H., 2022. Deepfake detection: A systematic literature review. *IEEE access*, *10*, pp.25494- 25513.

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

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

Rabhi, M., Bakiras, S. and Di Pietro, R., 2024. Audio-deepfake detection: Adversarial attacks and countermeasures. *Expert Systems with Applications*, *250*, p.123941.

Mubarak, R., Alsboui, T., Alshaikh, O., Inuwa-Dute, I., Khan, S. and Parkinson, S., 2023. A survey on the detection and impacts of deepfakes in visual, audio, and textual formats. *IEEE Access*.

Alshehri, A., Almalki, D., Alharbi, E. and Albaradei, S., 2024. Audio Deep Fake Detection with Sonic Sleuth Model. *Computers*, *13*(10), p.256.

Wang, C., Yi, J., Tao, J., Sun, H., Chen, X., Tian, Z., Ma, H., Fan, C. and Fu, R., 2022, October. Fully automated

end-to-end fake audio detection. In *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia* (pp. 27-33).

Conti, E., Salvi, D., Borrelli, C., Hosler, B., Bestagini, P., Antonacci, F., Sarti, A., Stamm, M.C. and Tubaro, S., 2022, May. Deepfake speech detection through emotion recognition: a semantic approach. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8962-8966). IEEE.

Huang, L. and Pun, C.M., 2024. Self-Attention and Hybrid Features for Replay and Deep-Fake Audio Detection. *arXiv preprint arXiv:2401.05614*.

Mary, A. and Edison, A., 2023, May. Deep fake Detection using deep learning techniques: A Literature Review. In *2023 International Conference on Control, Communication and Computing (ICCC)* (pp. 1-6). IEEE.

Chen, T., Kumar, A., Nagarsheth, P., Sivaraman, G. and Khoury, E., 2020, November. Generalization of Audio Deepfake Detection. In Odyssey (pp. 132-137).

Rana, M.S. and Sung, A.H., 2020, August. Deepfake stack: A deep ensemble-based learning technique for deepfake detection. In *2020 7th IEEE international conference on cyber security and cloud computing (CSCloud)/2020 6th IEEE international conference on edge computing and scalable cloud (EdgeCom)* (pp. 70-75). IEEE.

Patel, Y., Tanwar, S., Gupta, R., Bhattacharya, P., Davidson, I.E., Nyameko, R., Aluvala, S. and Vimal, V., 2023. Deepfake generation and detection: Case study and challenges. *IEEE Access*.

Yang, Y., Qin, H., Zhou, H., Wang, C., Guo, T., Han, K. and Wang, Y., 2024, April. A robust audio deepfake detection system via multi-view feature. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 13131-13135). IEEE.

Yang, W., Zhou, X., Chen, Z., Guo, B., Ba, Z., Xia, Z., Cao, X. and Ren, K., 2023. Avoid-df: Audio-visual joint learning for detecting deepfake. *IEEE Transactions on Information Forensics and Security*, *18*, pp.2015-2029.

Ali, M., Sabir, A. and Hassan, M., 2021, October. Fake audio detection using hierarchical representations learning and spectrogram features. In *2021 International Conference on Robotics and Automation in Industry (ICRAI)* (pp. 1-6). IEEE.

Wani, T.M. and Amerini, I., 2023, September. Deepfakes audio detection leveraging audio spectrogram and convolutional neural networks. In *International Conference on Image Analysis and Processing* (pp. 156- 167). Cham: Springer Nature Switzerland.

Almutairi, Z.M. and Elgibreen, H., 2023. Detecting fake audio of Arabic speakers using self-supervised deep learning. *IEEE Access*, *11*, pp.72134-72147.

Sivaramakrishnan, M., Rajput, A. and Saravanan, M., 2024, June. Classification of Deep Fake Audio Using MFCC Technique. In *2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS)* (pp. 1-6). IEEE.