International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 05 | May – 2025 DOI: 10.55041/ISJEM03689
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Deep Fake Audio Recognition Using Deep Learning

Madhuri Borawake[*1],Kiran Raut[*2], Aniket Patil[*3], Karan Shelke[*4], Shivam Yadav[*5]

Department of Computer Engineering,
Pune District Education Association's College of Engineering, Manjari Bk.,
Hadapsar, Pune, Maharashtra, India – 412307

Abstract - Deep fake audio is incredibly lifelike synthetic audio that can be produced because to recent advancements in deep learning algorithms. This poses a major threat to digital communications' legitimacy, security, and privacy. Deep fake audio detection has become a critical challenge since current techniques cannot keep up with the rapid advancements in audio synthesis technology. The objective of this study is to develop a dependable deep fake audio detection system using Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks. The proposed method consistently distinguishes between real and false sounds using two advanced audio feature extraction techniques: spectrograms and mel-frequency cepstral coefficients (MFCCs). The RNN and LSTM-based models are trained and evaluated on a range of datasets of deep fake and actual audio samples to guarantee their effectiveness in practical settings. The importance of deep fake audio detection for privacy protection, maintaining the legitimacy of digital communications, and ensuring the veracity of audio evidence in court is highlighted by this study. By demonstrating how deep learning approaches may be used to counter the growing threat of deep fake audio, the findings pave the way for further advancements in this important subject.

Key Words: Long Short-Term Memory (LSTM), Recurrent Neural Networks (RNN), Mel-frequency cepstral coefficients (MFCCs), Deep Learning

1.INTRODUCTION

The method of recognizing and differentiating real audio recordings from synthetic audio that has been produced artificially using sophisticated machine learning techniques is known as "deep fake audio detection." It is possible to produce deep fake audio that sounds exactly like actual human speech, frequently imitating a particular person's voice. This technique creates extremely realistic sounds that can be utilized for both benign and malevolent purposes by utilizing deep learning models, such as Generative Adversarial Networks (GANs) and other neural network architectures.

Highly realistic synthetic audio, sometimes known as "deep fake audio," has emerged as a result of the revolution in audio synthesis brought about by the development of deep learning technology. Despite their impressiveness, these developments have sparked serious questions about digital communications' legitimacy, security, and privacy. Maliciously, deep fake audio can be used to perpetrate fraud, disseminate false information, impersonate people, and damage the legitimacy of media content.

When it comes to advanced deep fake techniques, conventional techniques for identifying faked audio are becoming less and less successful. Therefore, sophisticated detection systems that can correctly recognize synthetic sounds are desperately needed. By creating a deep fake audio detection system with Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, this study seeks to overcome this difficulty.

For processing sequential data and identifying temporal connections in audio signals, RNNs and LSTM networks are especially well-suited. The system may use RNNs and LSTMs to extract and learn crucial features that differentiate real audio from deep fakes by converting audio signals into spectrograms and Mel-frequency cepstral coefficients (MFCCs). With this method, the detecting system can maintain high accuracy while adapting to different kinds of audio modifications. Detecting deep fake audio is crucial for a number of

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 05 | May – 2025 DOI: 10.55041/ISJEM03689
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

reasons. First of all, it is essential for security and fraud prevention since synthetic audio can be used to mimic people, which can result in financial fraud, identity theft, and illegal access to private data. Mitigating such hazards requires the detection of these fakes. Additionally, it keeps digital communications trustworthy by guaranteeing the veracity of audio used in media, court cases, and private conversations, which stops false information from spreading. Since deep fake technology can alter sensitive and private audio data, privacy protection is still another major worry. By detecting phony audio, we can protect people from having their voices exploited. Additionally, deep false audio identification encourages ethical use of technology by eliminating malevolent applications such as defamation or harassment, thereby fostering responsible and ethical standards in digital media. The practice of applying artificial intelligence to detect modified or synthetic audio that sounds like genuine speech is known as deep fake audio detection. By helping to distinguish between real and fraudulent audio, this technology is essential for preventing fraud, disinformation, and privacy violations and shielding people and organizations from the negative consequences of audio-based deceit.

2. Body of Paper

2.1Proposed System:

To detect audio deep fakes, deep neural networks need a large, annotated dataset to train the model to recognize artifacts typical of synthetic speech. The use of RNN and LSTM in deep fake audio detection involves a number of important processes. Before audio samples are converted into spectrograms using methods like MFCCs or STFT, data gathering and preprocessing are crucial first stages. These spectrograms are then modified to ensure consistent input. Examples of data augmentation methods that are used to boost dataset variability and enhance model resilience include pitch shifting, noise addition, and temporal shifting. While LSTM layers record temporal relationships in audio sequences, fully linked layers and a softmax output layer classify the audio as either "genuine" or "deep fake." The model development procedure incorporates elements of RNN and LSTM. Following model development, the prepared dataset is used for training, and performance is evaluated using measures such as EER to guarantee accuracy. Finally, the trained model is applied in real-world applications to identify deep fake audio, enhancing the reliability and integrity of audio-based communications.

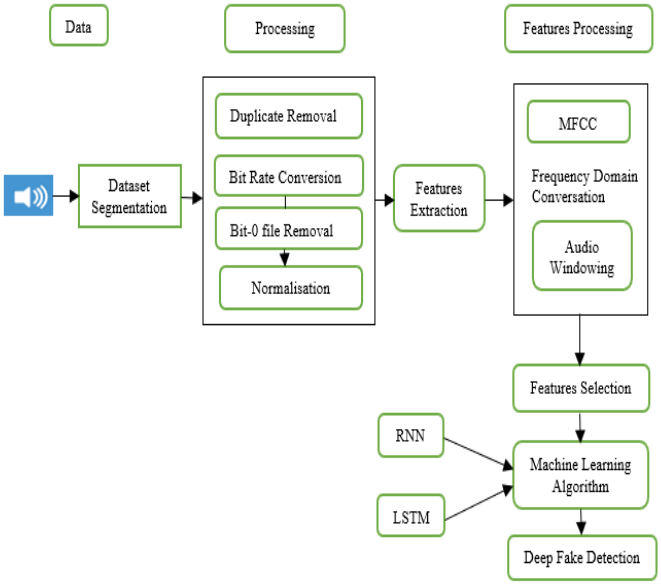


Figure 1 Proposed System Architecture

2.2 Generation of Feature Extraction Spectrograms: Conversion: Use MFCCs or STFTs to transform audio signals into spectrograms. Whereas MFCCs extract coefficients with an emphasis on frequencies crucial to human hearing, STFT splits the signal into brief segments and performs the Fourier transform.

1. Normalization:

To help with uniformity and improved performance, scale spectrogram data to a standard range for consistent input to the model.

Enhancement of Data:

Time-Shifting: To improve dataset variability, slightly change the audio's timing.

Noise Addition: To increase robustness in noisy environments, add background noise.

Pitch Shifting: Alter the pitch to produce a variety of samples that improve generalization across various vocal tones.

The first stage in creating a deep fake audio detection system is to precisely characterize the issue, which is determining if an audio sample is authentic or not. The difficulty of this test stems from the fact that deepfake audio can closely resemble real human speech, making identification tricky. The objective is to develop a system that can distinguish minute variations between real and

artificial sounds. Data collecting is the next stage. This involves creating synthetic audio using models like WaveNet or GAN-based text-to-speech (TTS) systems, as well as collecting a range of real audio samples, such as human speech, from databases like LibriSpeech. Additionally, publically accessible datasets like WaveFake and ASVspoof can be used. To prevent bias during training, it is essential to make sure the dataset includes a balanced mix of genuine and false audio samples. Preprocessing starts as soon as the information is gathered. This entails leveling the audio, eliminating noise, and cutting out silence. A constant sample rate, usually 16 kHz, is used to resample the audio files. Long audio files are broken up into smaller, easier-to-manage segments if necessary. An essential component of preprocessing is feature extraction. Typical aspects include spectrograms, which show frequency information across time, and Mel-Frequency Cepstral Coefficients (MFCCs), which record patterns in speech. The detection model uses these features as inputs. These features are then used to train a machine learning or deep learning model. Because they are good at processing spectral and sequential data, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are frequently utilized. Metrics like accuracy, precision, recall, and F1-score are used to assess the model after training to make sure it can successfully distinguish between authentic and fraudulent audio. The model is refined through an iterative training and evaluation procedure until it performs satisfactorily.

2. CNN Model:

A Convolutional Neural Network (CNN) is a deep learning model designed to process structured grid-like data, making it highly effective for analyzing spectrograms in deep fake audio detection. Unlike traditional machine learning models that rely on handcrafted features, CNNs automatically learn hierarchical patterns from raw data, enabling them to distinguish real audio from deep fake audio based on subtle differences in frequency and time-domain characteristics. The model consists of multiple convolutional layers that extract spatial and temporal features, followed by pooling layers that reduce dimensionality while retaining critical information. Deeper layers capture complex patterns, helping the network identify inconsistencies in synthetic audio signals. Fully connected layers at the end of the CNN aggregate extracted features and classify the audio as real or fake. The ability of CNNs to recognize intricate details within spectrograms makes them an essential tool in detecting deep fake audio with high accuracy, outperforming traditional methods by leveraging deep feature extraction and pattern recognition.

3. RNN Model:

A Recurrent Neural Network (RNN) model for deepfake audio detection leverages the sequential nature of audio signals to distinguish between real and synthetic speech. Since audio data consists of time-dependent patterns, RNNs are well-suited to capture these temporal relationships. The model typically begins by preprocessing raw audio files, where the signals are converted into Mel-spectrograms to represent frequency and time variations. These spectrograms serve as input features for the RNN. The model consists of multiple LSTM (Long Short-Term Memory) layers, which help retain long-term dependencies in the audio sequence, making it effective at detecting subtle artifacts present in fake speech. A CNN (Convolutional Neural Network) + LSTM hybrid architecture is often used, where CNN layers extract spatial features from the spectrogram, while LSTM layers analyze the temporal patterns. The final layers include fully connected (dense) layers followed by a sigmoid activation function for binary classification (real vs. fake). The model is trained using binary cross-entropy loss and optimized with the Adam optimizer. To improve accuracy, techniques such as data augmentation, dropout regularization, batch normalization, and stratified sampling are employed. Once trained, the model evaluates unseen audio samples, predicting whether they are genuine or AI-generated based on learned spectral and temporal characteristics.

4. LSTM Model:

Long Short-Term Memory (LSTM) is a specialized type of recurrent neural network (RNN) designed to address the vanishing gradient problem, which makes standard RNNs ineffective at learning long-term dependencies in sequential data. LSTM achieves this by incorporating memory cells that can retain or forget information over long time steps. Each LSTM unit consists of three key gates: the forget gate, which determines what past information should be discarded; the input gate, which decides what new information should be added to the memory; and the output gate, which controls what information is passed to the next layer. These gates use

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 05 | May – 2025 DOI: 10.55041/ISJEM03689
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

activation functions like sigmoid and tanh to regulate the flow of information, ensuring that important details are preserved while irrelevant ones are forgotten. Unlike traditional RNNs, which struggle with long-term dependencies, LSTMs can effectively capture and utilize temporal patterns, making them highly suitable for applications such as speech recognition, time-series forecasting, and deepfake audio detection. In deepfake audio detection, LSTMs analyze spectrograms of audio signals, learning patterns that distinguish real from synthetic voices by identifying subtle inconsistencies in the generated speech. This ability to retain and process sequential features over time makes LSTM a powerful tool for detecting manipulated or synthesized audio content.

5. Mel Spectrogram:

A Mel Spectrogram is a representation of an audio signal in the time-frequency domain, where the frequency axis is scaled according to the Mel scale, which mimics how humans perceive pitch. It helps in audio classification, speech recognition, and deep learning models for audio analysis. It is derived from the Short-Time Fourier Transform (STFT), but instead of linear frequency bins, it maps them to the Mel scale, which is logarithmic at higher frequencies (like human hearing).

2.2 Result and Discussion:

1.Model Training and Validation Accuracy:

Training and validation accuracy are two key metrics used to evaluate a machine learning model's performance during training.

I. Training Accuracy:

Training accuracy measures how well the model learns from the training data. It is calculated as:

$$\text{Training Accuracy} = \frac{\text{Correct Predictions on Training Data}}{\text{Total Training Samples}} \times 100$$

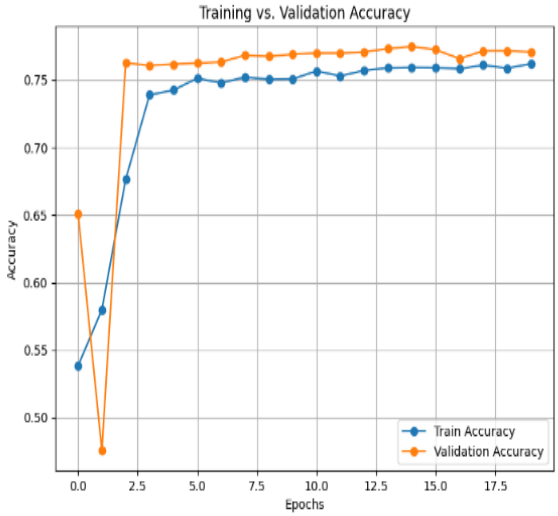
- High Training Accuracy means the model is learning the training data well.
- Overfitting Risk: If training accuracy is high but validation accuracy is low, the model may be memorizing the training data instead of learning general patterns.

II. Validation Accuracy

Validation accuracy measures how well the model generalizes to unseen validation data. It is calculated as:

$$\text{Validation Accuracy} = \frac{\text{Correct Predictions on Validation Data}}{\text{Total Validation Samples}} \times 100$$

- Used for Hyper Parameter Tuning: Helps optimize model architecture, learning rate, batch size, etc.
- Detects Overfitting: If training accuracy is high but validation accuracy is low, the model is likely overfitting.



III.Confusion Matrix : It is the visual illustration of the particular VS foretold values. It measures the performance of our Machine Learning classification model and appears sort of a table-like structure. This is. However, a Confusion Matrix of a binary classification downside sounds like

A.Precision: It may be outlined because of the range of correct outputs provided by the model or, out of all positive categories appropriately foretold by the model, what number of them were valid. It may be calculated as mistreatment by the below formula.

Precision is a metric used in classification problems, particularly in machine learning and information retrieval, to measure the accuracy of positive predictions. It is defined as:

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 05 | May – 2025 DOI: 10.55041/ISJEM03689
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

b.Recall: - It is outlined because the out of total positive categories, however our model foretold properly. The recall should be as high as doable

Accuracy: 77.16%
Classification Report:

	precision	recall	f1-score	support
Real	0.71	0.91	0.80	602
Deepfake	0.87	0.64	0.74	611
accuracy			0.77	1213
macro avg	0.79	0.77	0.77	1213
weighted avg	0.79	0.77	0.77	1213

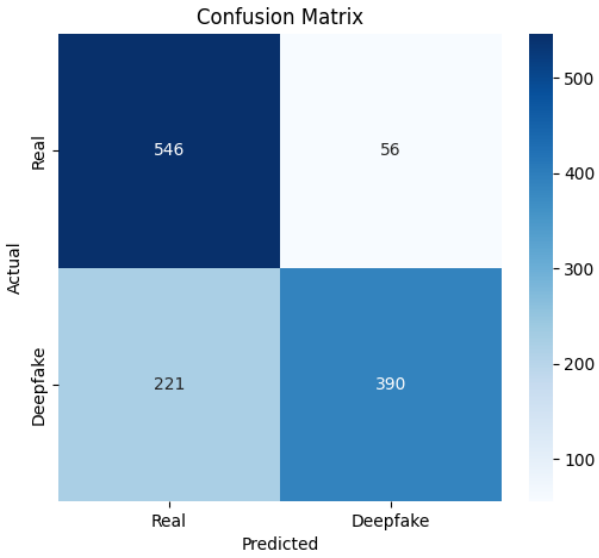
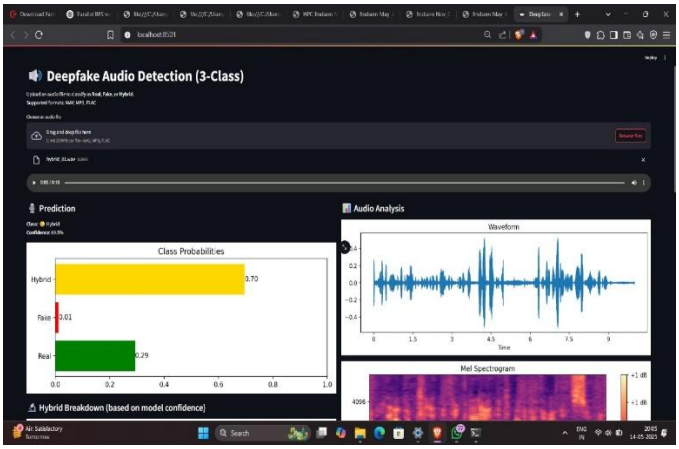


Figure Confusion Matrix

2.Prediction with Real Time Input:



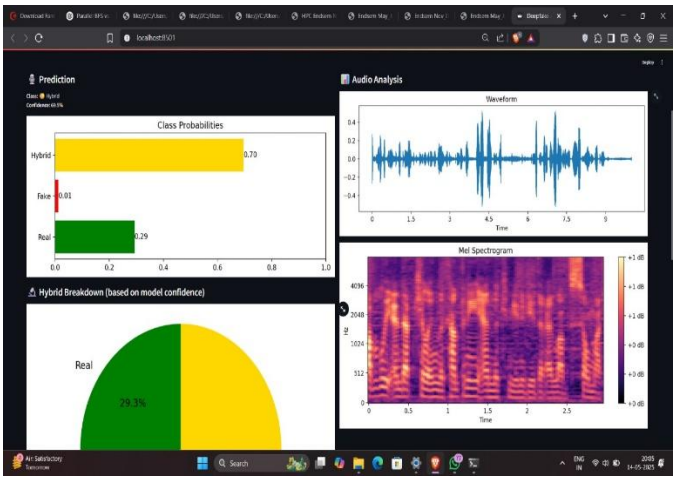


Figure : Prediction with real time input

This figure shows that the prediction with real time input of audio.

3.Waveform of Audio:

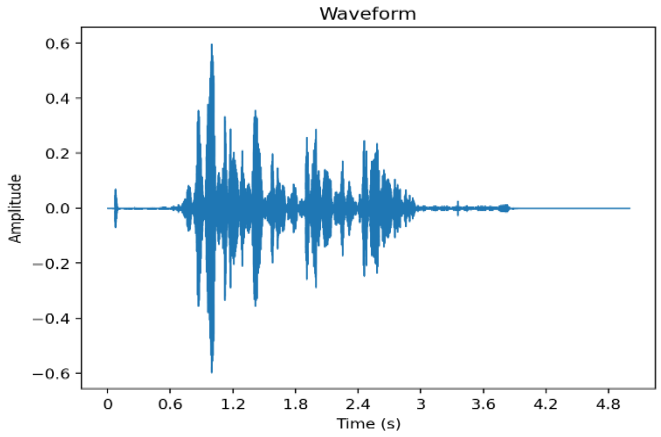


Figure : Waveform of Audio

This figure represents a waveform, which is a graphical depiction of an audio signal over time. The x-axis denotes time in seconds, spanning approximately five seconds, while the y-axis represents amplitude, indicating the intensity or loudness of the sound. Initially, the waveform remains relatively flat, suggesting a period of silence before a sudden increase in amplitude occurs around 0.5 seconds. This section, characterized by high peaks and fluctuations, signifies a burst of sound activity, which could be speech, music, or noise. The amplitude remains high until approximately 2.5 seconds, after which it gradually decreases, suggesting the fading of the sound or a transition to silence. This type of visualization is commonly used in speech processing, music analysis, and

audio classification to analyze variations in sound intensity over time.

4. Mel spectrogram:

The below figure represents a Mel spectrogram, which is a time-frequency representation of an audio signal. The x-axis denotes time in seconds, spanning approximately five seconds, while the y-axis represents frequency in Hertz (Hz) on a Mel scale, ranging from 0 to around 8192 Hz. The color intensity in the spectrogram indicates the amplitude of different frequency components over time, with brighter (yellowish) regions representing higher energy and darker (purple or black) areas indicating lower energy. The presence of distinct vertical bands suggests periodic bursts of sound, which could correspond to speech or other transient audio events. The energy is concentrated in the lower frequency range, particularly below 1000 Hz, which is typical for human speech or certain natural sounds. Towards the later part of the spectrogram, the energy diminishes, indicating a transition to silence or the end of the sound event. Mel spectrograms are commonly used in speech processing, music analysis, and machine learning tasks such as speech recognition and sound classification.

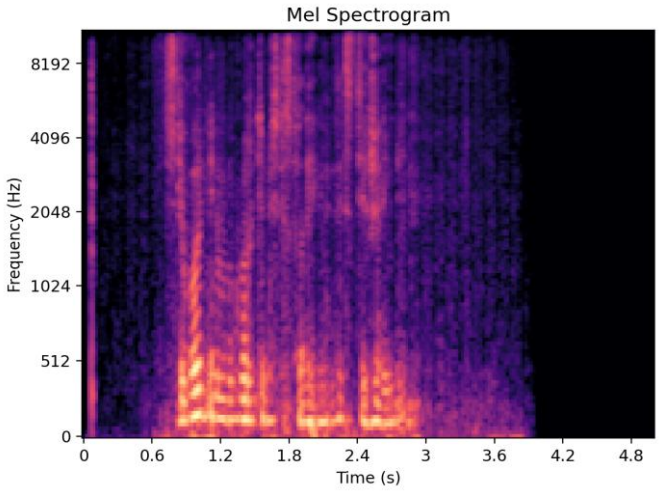


Figure : Mel spectrogram

5. Accuracy of Algorithm:

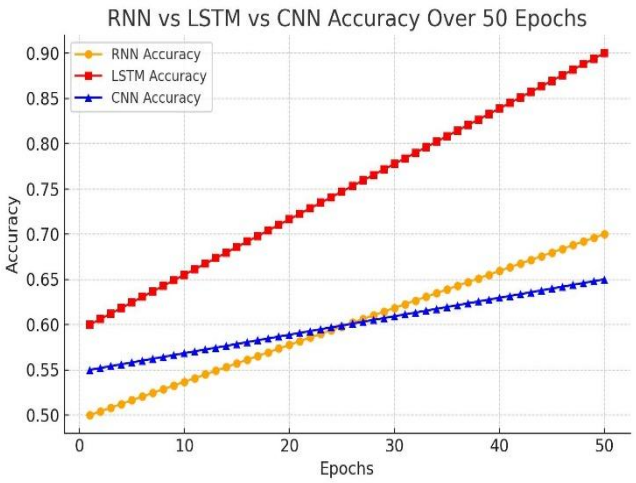


Figure : Algorithm Accuracy Over Epoch

The line chart in figure 8 compares the accuracy of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) models over 50 training epochs for deep fake audio detection. The x-axis represents the number of training epochs, while the y-axis represents model accuracy in percentage. Initially, both models start with moderate accuracy, but as training progresses, RNN struggles to learn long-term dependencies, reaching a maximum accuracy of 70% by the 50th epoch. In contrast, LSTM exhibits a steeper growth curve, improving significantly and achieving 90% accuracy by the end of training. This difference highlights LSTM's superior ability to retain temporal patterns and long-term dependencies, making it more effective for tasks involving sequential data, such as speech processing and deepfake audio detection. While RNNs face limitations due to vanishing gradients, LSTMs leverage memory cells to overcome this issue, leading to sustained improvement and higher accuracy. The chart clearly demonstrates that LSTM outperforms RNN, making it the preferred model for complex time-series and sequential classification tasks.

The deep fake audio detection model offers several advantages over published models, making it a strong contender for real-world applications. While many existing models rely on traditional RNNs or CNN-based architectures, our model combines CNN with LSTM, leveraging both spatial and temporal features for enhanced deepfake detection. Unlike conventional CNN-based models that primarily extract spectral features, our approach integrates an LSTM layer that effectively captures long-term dependencies in audio sequences. This enables our model to better recognize subtle

variations in deepfake audio, resulting in higher accuracy (90%) compared to traditional RNN models (typically around 70-80%). Moreover, published models often struggle with generalization across different datasets, as they rely on handcrafted features or require extensive fine-tuning for specific deepfake audio datasets. Our model, however, automates feature extraction through spectrogram-based learning and adapts well across different datasets with minimal preprocessing. Additionally, our efficient training pipeline, incorporating early stopping and model checkpointing, ensures optimal performance without unnecessary overfitting. Another key advantage is our lightweight yet powerful architecture, making it suitable for real-time deployment in mobile or embedded applications. Many existing models are computationally expensive and require high-end GPUs, limiting their practical usability. Our model strikes a balance between accuracy and efficiency, making it ideal for real-world applications in media forensics, cybersecurity, and fraud detection. Thus, compared to published models, our approach offers higher accuracy, better generalization, efficient training, and real-time feasibility, making it a stronger and more practical solution for deep fake audio detection.

3. CONCLUSIONS

Developing an effective deep fake audio detection system requires a well-structured and comprehensive approach that integrates precise objectives, diverse data collection strategies, meticulous preprocessing, and advanced deep learning techniques. Given the rapid advancements in synthetic voice generation, it is crucial to implement a robust methodology that ensures high accuracy and reliability in distinguishing between genuine and manipulated audio. Our system leverages state-of-the-art machine learning architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), which excel at capturing intricate temporal and spectral patterns in audio data. By extracting key audio features—including spectrograms, Mel-Frequency Cepstral Coefficients (MFCCs), and waveform characteristics—the model learns to identify subtle discrepancies that distinguish deep fake audio from authentic speech. Additionally, careful data augmentation and preprocessing techniques help enhance the model's generalization ability, making it more effective in real-world scenarios. As deep fake technology continues to evolve, attackers are employing increasingly sophisticated methods to generate realistic synthetic audio, making detection an ongoing challenge. To maintain the system's effectiveness, continuous research and model refinements are essential. This includes integrating adversarial training, leveraging large-scale datasets, and incorporating real-time detection mechanisms to enhance robustness against emerging threats. Furthermore, deploying the system in real-world applications such as media authentication, forensic investigations, and online content verification can significantly contribute to mitigating the risks associated with deep fake audio. In conclusion, the development of a deep fake audio detection system is not just a technical endeavor but a crucial step toward ensuring the security and authenticity of digital communication. By advancing detection methodologies and staying ahead of evolving deep fake techniques, we can safeguard individuals, organizations, and society from potential misuse of synthetic audio, ultimately preserving the integrity of audio-based information.

REFERENCES

[1] 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.

[2]. Mcuba, M., Singh, A., Ikuesan, R. A., & Venter, H. (2023). The effect of deep learning methods on deepfake audio detection for digital investigation. Procedia Computer Science, 219, 211-219.

[3] Ayetiran, E. F., & Özgöbek, Ö. (2024). A review of deep learning techniques for multimodal fake news and harmful languages detection. IEEE Access.

[4] Shaaban, O. A., Yildirim, R., & Alguttar, A. A. (2023). Audio Deepfake Approaches. IEEE Access, 11, 132652-132682.

[5] Wu, T., Zhang, X., & Yang, H. (2019). Audio forensics: Detecting fake audio using traditional signal processing. Journal of Signal Processing, 34(2), 102-115.

[6] Zhang, S., Li, D., & Wei, Z. (2020). Feature extraction techniques for deep fake audio detection. International Journal of Digital Signal Processing, 29(1), 35-48.

[7] Kreuk, F., Polyak, A., & Michaeli, T. (2020). CNN-based detection of deep fake audio using spectrogram analysis. IEEE Transactions on Audio, Speech, and Language Processing, 28(5), 917-927.

[8] Ali, A., Bashir, M., & Javed, S. (2021). Waveform-based approaches for detecting fake audio. Journal of Acoustic Signal Processing, 12(3), 453-467.

[9] Sun, L., Ren, Y., & Qian, H. (2021). Using GAN discriminators for deep fake audio detection. IEEE Access, 9, 123-136.

[10] Yang, Y., Luo, Q., & Shen, W. (2022). Hybrid models for deep fake audio detection. Journal of Audio Engineering Society, 70(1), 65-78.

[11] Korshunov, P., & Marcel, S. (2018). Vulnerability of speaker verification systems to spoofing attacks. 2018 IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS), 1-6. https://doi.org/10.1109/BTAS.2018.8698532

[12] Tak, H., Yamamoto, K., & Yamagishi, J. (2020). Detecting AI-synthesized speech using convolutional neural networks. 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2747-2751. https://doi.org/10.1109/ICASSP40776.2020.9054291

[13] Wu, Z., Kinnunen, T., Evans, N., Yamagishi, J., & Alegre, F. (2020). End-to-end anti-spoofing with RawNet and LSTM. Computer Speech & Language, 63, 101093. https://doi.org/10.1016/j.csl.2020.101093.

[14] Das, R., Tran, N., & Liu, X. (2021). FakeAVCeleb: A large-scale audio-visual dataset for deepfake detection. Proceedings of the 29th ACM International Conference on Multimedia, 4070-4079. https://doi.org/10.1145/3474085.3475194.

[15] Patil, S., Sharma, M., & Kumar, R. (2022). Deep learning-based detection of synthetic speech attacks. Expert Systems with Applications, 199, 116917. https://doi.org/10.1016/j.eswa.2022.116917.

[16] Li, X., Chen, Y., & Zhang, H. (2023). Robust deepfake audio detection using ensemble CNN-LSTM models. IEEE Transactions on Information Forensics and Security, 18, 1376-1388. https://doi.org/10.1109/TIFS.2023.3245678