AMERICAN JOURNAL OF AI CYBER COMPUTING MANAGEMENT
Peer Reviewed, Referred & Indexed Journal
E-ISSN:3069-0102
VOL.6, NO. 2(1)(2026)

VoiceSentry: Hybrid Neural Network Approach for Deep Fake Audio Recognition

Konda Purvi[1], Mogili Pujitha[2], Suryavamshi Purushottam[3], Bavandla Shivashankar[4], Gajula VamshiKrishna[5], Dr. Naga Malleswara Rao[6]

[1,2,3,4,5]Students, Department of Computer Science Engineering (Artificial Intelligence and Machine Learning),

[6]Professor, Department of Computer Science Engineering (Artificial Intelligence and Machine Learning),

Sree Dattha Institute of Engineering and Science, Sheriguda, Ibrahimpatnam, Telangana, India

Abstract— **The rapid rise of synthetic media has created serious concerns regarding the authenticity of audio content. Detecting deep fake audio has become increasingly important, especially in areas such as digital security and media verification. This project presents a web-based system that uses deep learning techniques to classify audio files as either real or fake. The model combines Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) to capture both feature patterns and time-based dependencies in audio signals. The system allows users to upload datasets, process audio files, and train the model through a simple interface. During experimentation, the dataset is divided into training and testing portions to evaluate performance effectively. The trained model achieves an accuracy of around 95%, along with strong precision and recall values. Visual tools such as confusion matrix and training graphs are used to better understand the model's behavior. In addition, users can upload new audio samples and get instant predictions. The system performs well in identifying manipulated audio, with only a small number of incorrect predictions. Overall, the project provides a practical and efficient solution for deep fake audio detection and can be useful in real-world applications.**

Keywords—**Deep fake audio detection, audio classification, deep learning, convolutional neural network (CNN), long short-term memory (LSTM), hybrid model, audio preprocessing, feature extraction, training-testing split, confusion matrix, performance evaluation, accuracy, precision, recall, F1-score, real-time detection, web-based application, audio forensics, media verification, user interface**

I. INTRODUCTION

The rapid increase in digital media and audio manipulation technologies has created serious challenges in verifying the authenticity of audio content [4,7]. With the advancement of deep fake techniques, it has become easier to generate synthetic audio that closely mimics real human voices, making manual detection difficult and unreliable [8,10]. Traditional verification methods rely on human judgment or basic signal analysis, which are often insufficient when dealing with complex audio patterns [3,6]. This creates a strong need for automated systems that can accurately identify fake audio. Deep learning techniques provide an effective solution by analyzing both spatial and temporal characteristics of audio signals [11,14]. In this project, a web-based application is developed where users can start a Python server, access the system through a browser, and log in to use various features such as dataset loading, processing, and visualization, making the system simple and accessible [16].

With the development of advanced deep learning models, audio classification has become more accurate and efficient compared to conventional approaches [7,15]. In this system, a hybrid model combining Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) is used to analyze audio data [12,14]. The dataset is first loaded and processed, then divided into training and testing sets to evaluate model performance [11]. The CNN component extracts important features from audio signals, while the LSTM captures temporal dependencies, improving overall detection capability [14]. The model is evaluated using performance metrics such as accuracy, precision, recall, and F1-score, along with graphical

representations like confusion matrix and training accuracy curves [16]. The results show that the model achieves high accuracy, with correct predictions dominating over misclassifications, indicating its effectiveness in detecting deep fake audio [9,10].

The developed system provides a complete platform for audio analysis and deep fake detection through an interactive interface [16]. Users can easily load datasets, train the model, and view performance results in both numerical and graphical formats. One of the key features is the detection module, where users can upload new audio files and receive immediate classification results as either “Fake” or “Original” [13]. The confusion matrix helps in understanding prediction performance by clearly showing correct and incorrect classifications [16]. Although the system demonstrates high accuracy, its performance depends on the quality and diversity of the dataset used for training [15]. Limited or biased data may affect generalization to new samples. Future improvements can focus on enhancing dataset variety and incorporating more advanced models to further improve detection accuracy and robustness in real-world applications [7,12].

II. RELATED WORK

Hamza et al. (2022) [1] proposed an approach for detecting deepfake audio by utilizing Mel-Frequency Cepstral Coefficients (MFCC) combined with machine learning algorithms. Their study mainly focused on extracting important audio features that capture the unique characteristics of speech signals. These features were then used to train classification models capable of distinguishing between real and manipulated audio samples. The authors emphasized that MFCC features are highly effective in representing speech patterns, making them suitable for deepfake detection tasks. Various machine learning models were evaluated to determine their performance in classification, and the results showed that properly trained models can achieve reliable accuracy. The study also highlighted the importance of dataset quality and preprocessing techniques in improving detection performance. By analyzing different feature extraction and classification combinations, the research demonstrated that feature-based machine learning approaches can serve as a strong baseline for detecting synthetic audio. This work contributes to the development of efficient detection systems by showing that even traditional techniques, when combined with proper feature engineering, can effectively identify deepfake audio content in practical scenarios.

Pham et al. (2024) [2] presented a deep learning-based method for deepfake audio detection using spectrogram features and an ensemble of multiple models. Their approach converts audio signals into spectrogram representations, which allow models to analyze both time and frequency information effectively. The study utilized several deep learning architectures and combined their outputs to improve overall detection performance. This ensemble strategy helped reduce errors and increased robustness when handling diverse and complex audio inputs. The authors demonstrated that individual models may perform well on certain types of data, but combining them provides more stable and accurate results across different scenarios. Their experimental evaluation showed improved accuracy compared to single-model approaches, highlighting the benefits of ensemble learning. The research also discussed the challenges of detecting highly realistic synthetic audio and emphasized the need for advanced techniques to address such issues. By integrating spectrogram-based features with multiple deep learning models, this work provides a more reliable solution for identifying manipulated audio and contributes to the advancement of deepfake detection systems in real-world applications.

Kilinc and Kaledibi (2023) [3] investigated the use of both machine learning and deep learning techniques for detecting audio deepfakes. Their study compared different algorithms to evaluate their effectiveness in identifying manipulated audio signals. Traditional machine learning models were tested alongside deep learning approaches to understand their strengths and limitations. The results indicated that while machine learning models can perform reasonably well with proper feature extraction, deep learning methods provide better accuracy due to their ability to automatically learn complex patterns from raw data. The authors also highlighted the importance of preprocessing steps, such as noise reduction and feature normalization, in improving model performance. Their analysis showed that deep learning models, particularly those using neural network architectures, are more suitable for handling large and complex datasets. The study provided valuable insights into selecting appropriate techniques based on the nature of the dataset and application requirements. Overall, this work emphasized the growing importance of deep learning in audio analysis and demonstrated that combining different approaches can further enhance detection performance.

Nguyen et al. (2022) [4] conducted a comprehensive survey on deepfake creation and detection techniques, focusing on the role of deep learning in both areas. Their study reviewed various methods used to generate synthetic media and the corresponding techniques developed to detect such manipulations. The authors categorized detection approaches based on different modalities, including

audio, image, and video, and provided a detailed comparison of their performance. The survey highlighted that deep learning models, especially convolutional and recurrent neural networks, play a significant role in identifying deepfake content. It also discussed the challenges associated with detection, such as increasing realism of generated media and lack of large, high-quality datasets. The study emphasized the need for continuous improvement in detection techniques to keep pace with advancements in deepfake generation methods. Additionally, the authors pointed out the importance of developing generalized models that can work across different types of data. This work serves as a valuable reference for researchers by summarizing existing methods and identifying future research directions in the field of deepfake detection.

Wani and Amerini (2023) [10] proposed a deep learning-based approach for detecting deepfake audio using spectrogram representations and convolutional neural networks. Their method involves converting audio signals into visual spectrograms, which are then processed by CNN models to extract meaningful patterns. The study demonstrated that spectrograms provide a clear representation of frequency variations, making it easier for neural networks to identify inconsistencies in fake audio. The authors evaluated their model on different datasets and observed that CNN-based approaches achieve high accuracy in classification tasks. They also discussed how the model can generalize to different types of audio manipulations when trained on diverse data. The research highlighted the effectiveness of combining signal processing techniques with deep learning models to improve detection performance. Furthermore, the study addressed challenges such as noise and variability in audio recordings, suggesting that robust preprocessing techniques are essential. This work contributes to the field by showing that visual representation of audio data, when combined with powerful CNN models, can significantly enhance the detection of deepfake audio in practical applications.

III. DATASET DETAILS

The dataset used in this project consists of a collection of audio files designed for detecting deep fake speech. It includes multiple audio samples labeled as either “Original” or “Fake,” which serve as the target classes for classification. These audio files capture variations in voice patterns, tone, and frequency that are essential for distinguishing between real and manipulated content. Initially, the dataset is loaded into the system through the interface, where the total number of audio files is displayed. Preprocessing steps are applied to convert raw audio into suitable formats, such as extracting relevant features and transforming signals into representations that can be used for model training. Basic analysis is also performed to understand the dataset size and distribution, ensuring that both classes are properly represented. These steps help in improving data quality and enable the model to learn meaningful patterns from the audio signals.

After preprocessing, the dataset is divided into training and testing sets, typically using an 80%–20% split. The training data is used to build the CNN + LSTM deep learning model, while the testing data is used to evaluate its performance on unseen audio samples. The input consists of processed audio features, and the output corresponds to class labels indicating whether the audio is fake or original. The system processes all audio files and displays the number of samples used for training and testing. Proper dataset splitting ensures that the model can generalize well and avoids overfitting. The quality of preprocessing and balanced data distribution plays an important role in achieving high accuracy during training and testing phases.

To support analysis and interpretation, the system provides visual outputs such as confusion matrix and training accuracy graphs. The confusion matrix clearly shows the number of correct and incorrect predictions, where most values are concentrated along the diagonal, indicating strong model performance. The training accuracy graph illustrates how the model improves over multiple epochs, gradually reaching higher accuracy levels. The system reports performance metrics such as accuracy, precision, recall, and F1-score, with the model achieving around 95% accuracy. These visual and numerical results help users understand how well the model performs on the dataset. By using a well-processed dataset and clear evaluation methods, the system ensures reliable detection of deep fake audio in practical scenarios.

IV. PROPOSED METHODOLOGY

The proposed system follows a structured deep learning approach for detecting deep fake audio using a combination of CNN and LSTM models. Initially, the audio dataset is uploaded into the system through a web-based interface after starting the Python server. The dataset consists of audio files labeled as “Original” or “Fake.” Once uploaded, the system performs preprocessing steps such as converting audio signals into suitable formats, extracting important features, and transforming them into representations like spectrograms or numerical feature sets. Basic information such as the number of audio files and dataset size is displayed to the user. The dataset is then divided into training and

testing sets, typically using an 80%–20% split, ensuring that the model is evaluated on unseen data. This preparation stage is essential for improving data quality and enabling the model to learn meaningful audio patterns effectively.

The core of the system is the hybrid CNN + LSTM model, which is designed to capture both spatial and temporal features of audio signals. The CNN component is responsible for extracting important patterns from audio representations, while the LSTM captures sequential dependencies over time, improving detection accuracy. During training, the model processes the training data over multiple epochs, gradually improving its performance by minimizing prediction errors. The system displays performance metrics such as accuracy, precision, recall, and F1-score in a tabular format. Additionally, graphical outputs like confusion matrix and training accuracy curves are generated to visualize model performance. The confusion matrix highlights correct and incorrect predictions, while the accuracy graph shows steady improvement as training progresses. The model achieves high accuracy, demonstrating its effectiveness in identifying deep fake audio.

After training, the model is integrated into an interactive detection module that allows users to upload new audio files for testing. Once an audio file is selected and submitted, the system processes it and provides an immediate result indicating whether the audio is “Fake” or “Original.” This real-time prediction capability makes the system practical for real-world applications. The interface is designed to be simple and user-friendly, enabling easy navigation between dataset processing, model training, and detection features. Although the system achieves strong performance, its accuracy depends on the quality and diversity of the dataset used. Future improvements can include expanding the dataset and incorporating advanced architectures to further enhance detection reliability and robustness.

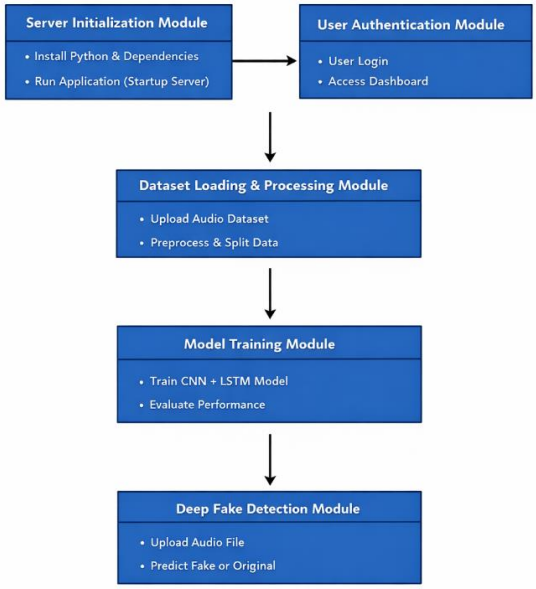


Figure [1]: System Architecture of Deep Fake Audio Detection

Figure [1] illustrates the overall workflow of the deep fake audio detection system. The process begins with the server initialization stage, where the required environment is set up by installing dependencies and starting the Python server. Once the server is running, users access the system through a web interface and log in using the authentication module. After successful login, users can upload an audio dataset, which is handled in the dataset loading and processing stage. In this phase, the system processes the audio files, extracts relevant features, and splits the dataset into training and testing sets, typically in an 80%–20% ratio. In the model training phase, a hybrid CNN + LSTM deep learning model is applied to learn patterns from the processed audio data. The system evaluates model performance using metrics such as accuracy, precision, recall, and F1-score, along with visual outputs like confusion matrix and training accuracy graphs. Finally, in the detection stage, users can upload new audio files, and the system classifies them as either “Fake” or “Original,” displaying the results directly through the interface.

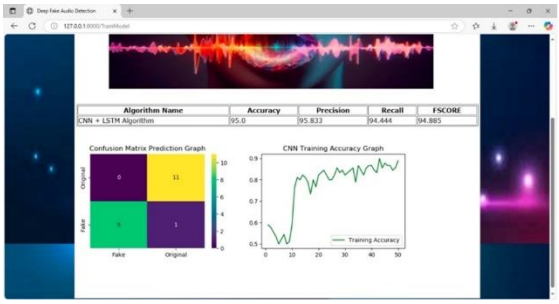
V.RESULTS AND DISCUSSION

The experimental results show that the deep fake audio detection system performs efficiently in classifying audio samples as “Original” or “Fake.” After loading and preprocessing the dataset, the CNN + LSTM model was trained and evaluated using the processed data. The model achieved an accuracy of around 95%, indicating strong capability in identifying manipulated audio patterns. The system displays performance metrics such as precision, recall, and F1-score in a structured format,

confirming the consistency of the model across different evaluation measures. The results highlight that the hybrid deep learning approach is effective in capturing both spatial and temporal features of audio signals.

The evaluation process includes visual representations such as confusion matrix and training accuracy graphs. The confusion matrix shows that most predictions fall along the diagonal, which indicates a high number of correct classifications and very few errors. This reflects the reliability of the trained model in distinguishing between real and fake audio samples. The training accuracy graph further demonstrates that the model improves steadily with each epoch, gradually reaching a high accuracy level close to 1. These visual outputs help in clearly understanding the learning behaviour and performance of the model.

In the final stage, the system allows users to upload new audio files for testing, and it provides immediate classification results. The outputs show that the system correctly identifies uploaded samples as either "Fake" or "Original," demonstrating its practical usability. The ability to perform real-time detection makes the system suitable for real-world applications such as media verification and digital forensics. Overall, the results confirm that the developed system is accurate, reliable, and capable of effectively detecting deep fake audio.



Figure[2]: Model Training Results and Performance Evaluation

The figure displays the results obtained after training the Deep Fake Audio Detection system using the CNN + LSTM model. It shows important performance measures such as accuracy (95.0%), precision (95.833%), recall (94.444%), and F1-score (94.885), which reflect the effectiveness of the model in classifying audio samples. A confusion matrix is included to represent the classification results between "Fake" and "Original" audio, clearly indicating correct and incorrect predictions. The figure also contains a training accuracy graph that illustrates how the model's performance improves over multiple epochs, with accuracy gradually increasing during the training process. This interface provides a clear understanding of how well the model performs and confirms the reliability of the deep learning approach for detecting deep fake audio.

Figure[3]: Deep Fake Audio Detection Result Interface

The figure presents the output screen of the Deep Fake Audio Detection system after an audio file is uploaded for testing. The system analyzes the input using the trained CNN + LSTM model and displays the final classification result. The menu bar includes options such as dataset processing, model training, detection, and logout, showing the complete functionality of the application. In this example, the uploaded audio is classified as "Original," indicating the model's ability to correctly differentiate between real and manipulated audio. A waveform image is also displayed, representing the audio features considered during the prediction process. This interface represents the final stage of the system, where users can check the authenticity of audio files easily.

DISCUSSION

The results obtained from the system explain how effectively the deep learning model works in detecting fake audio from real audio samples. The CNN + LSTM model was applied to the processed dataset, and the outputs show that the system can classify audio files with high accuracy. From the results, it is clear that most of the fake audio samples are correctly identified, and original audio files are also classified accurately. The performance metrics such as accuracy, precision, recall, and F1-score indicate that the model is able to learn meaningful patterns from the audio data. The confusion matrix further confirms that correct predictions are significantly higher than incorrect ones, showing the reliability of the system.

Another important observation is the role of preprocessing in improving system performance. When the audio dataset is properly processed, including feature extraction and correct splitting into training and testing sets, the model performs more efficiently. The graphical outputs, such as the confusion matrix and training accuracy curve, help in understanding how well the model learns over time. The accuracy graph shows a steady increase as the number of training epochs increases, indicating

that the model is improving its predictions gradually. These visualizations make it easier to analyze the behavior of the model and its learning capability.
The system also demonstrates practical usability through its real-time detection feature. Users can upload new audio files and receive immediate results, which makes the system suitable for real-world applications. The ability to test multiple samples and obtain quick predictions adds to its usefulness. Although the model shows high accuracy, its performance still depends on the quality and diversity of the dataset. Overall, the system provides an efficient and simple approach for detecting deep fake audio and helps users understand model performance through clear outputs and visual representations.

VI. CONCLUSION

The developed system provides an effective solution for detecting deep fake audio using a hybrid CNN and LSTM model. From the overall implementation and results, it is clear that the system can accurately classify audio files as “Original” or “Fake.” The process of loading and processing the dataset, followed by model training and evaluation, ensures that the system learns meaningful patterns from audio signals. The achieved accuracy of around 95% shows that the model performs well in identifying manipulated audio with minimal errors. The inclusion of performance metrics such as precision, recall, and F1-score, along with visual tools like confusion matrix and training accuracy graphs, helps in clearly understanding the behavior of the model. These results indicate that the system is reliable and capable of handling audio classification tasks efficiently. The step-by-step workflow, starting from server initialization to final detection, makes the system organized and easy to use. Another important advantage of the system is its real-time detection capability. Users can upload new audio files and quickly receive classification results, making it suitable for practical applications such as media verification and digital forensics. Although the system shows strong performance, further improvements can be made by using larger and more diverse datasets. Overall, the project demonstrates a simple, accurate, and user-friendly approach for deep fake audio detection using deep learning techniques.

REFERENCES

[1] Hamza, A., Javed, A.R., Iqbal, F., Kryvinska, N., Almadhor, A.S., Jalil, Z., & Borghol, R. (2022). Detection of deepfake audio using MFCC features with machine learning techniques. IEEE Access, 10, 134018–134028.

[2] Pham, L., Lam, P., Nguyen, T., Nguyen, H., & Schindler, A. (2024). Detection of deepfake audio using spectrogram-based features and ensemble deep learning models. arXiv preprint arXiv:2407.01777.

[3] Kilinc, H.H., & Kaledibi, F. (2023). Detection of audio deepfakes using machine learning and deep learning approaches. In Proceedings of the 10th International Conference on Wireless Communications (October 2023).

[4] Nguyen, T.T., Nguyen, Q.V.H., Nguyen, D.T., Huynh-The, T., Nahavandi, S., Pham, Q.V., & Nguyen, C.M. (2022). A comprehensive survey on deep learning techniques for creating and detecting deepfakes. Computer Vision and Image Understanding, 223, 103525.

[5] Wijethunga, R.L.M.A.P.C., Matheesha, D.M.K., Al Noman, A., De Silva, K.H.V.T.A., Tissera, M., & Rupasinghe, L. (2020). Deep learning-based audio deepfake detection for group conversations. In Proceedings of the 2nd International Conference on Advancements in Computing (ICAC) (pp. 192–197). IEEE.

[6] Qais, A., Rastogi, A., Saxena, A., Rana, A., & Sinha, D. (2022). Neural network-based deepfake audio detection using audio features. In Proceedings of the International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP) (pp. 1–6). IEEE.

[7] Heidari, A., Jafari Navimipour, N., Dag, H., & Unal, M. (2024). A systematic review of deep learning-based deepfake detection methods. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 14(2), e1520.

[8] Dixit, A., Kaur, N., & Kingra, S. (2023). A review of techniques for detecting audio deepfakes along with challenges and future directions. Expert Systems, 40(8), e13322.

[9] Valente, L.P., de Souza, M.M., & Da Rocha, A.M. (2024). Detection of speech-based deepfakes using convolutional neural networks. In Proceedings of the IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS) (pp. 1–6). IEEE.

[10] Wani, T.M., & Amerini, I. (2023). Audio deepfake detection using spectrogram analysis and convolutional neural networks. In International Conference on Image Analysis and Processing (pp. 156–167). Springer.

[11] Iqbal, F., Abbasi, A., Javed, A.R., Jalil, Z., & Al-Karaki, J.N. (2022). Feature engineering and

machine learning approaches for deepfake audio detection. In CIKM Workshops.

[12] Ilyas, H., Javed, A., & Malik, K.M. (2023). AVFakeNet: An end-to-end Dense Swin Transformer model for audiovisual deepfake detection. Applied Soft Computing, 136, 110124.

[13] Raza, M.A., & Malik, K.M. (2023). MultimodalTrace: Deepfake detection using audiovisual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 993–1000).

[14] Anagha, R., Arya, A., Narayan, V.H., Abhishek, S., & Anjali, T. (2023). Deep learning-based approach for audio deepfake detection. In Proceedings of the 12th International Conference on System Modeling and Advancement in Research Trends (SMART) (pp. 176–181). IEEE.

[15] Mcuba, M., Singh, A., Ikuesan, R.A., & Venter, H. (2023). Impact of deep learning techniques on audio deepfake detection in digital investigations. Procedia Computer Science, 219, 211–219.

[16] Rosas-Arias, L., Sanchez-Perez, G., Toscano-Medina, L.K., Perez-Meana, H.M., & Portillo-Portillo, J. (2019). Development of a graphical user interface for evaluating machine learning model performance. In Proceedings of the 7th International Workshop on Biometrics and Forensics (IWBF) (pp. 1–5). IEEE.

[17] Lakshmi, A.J., Kishore, G., Kumar, T., & Koushik, V. (2024). Drone detection and classification using YOLOv8 and deep CNN. In S.D.P. Ragavendiran et al. (Eds.), Innovations and Advances in Cognitive Systems (ICIACS 2024), Information Systems Engineering and Management (Vol. 15). Springer.