IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

DEEPFAKE AUDIO DETECTION MODEL BASED ON MEL SPECTROGRAM USING CONVOLUTIONAL NEURAL NETWORK

[1]Fathima G, [2]Kiruthika S, [3]Malar M, [4]Nivethini T

[1]Professor, [2]Student, [3]Student, [4]Student

[1]Department of Computer Science and Engineering,

Adhiyamaan College of Engineering, Hosur, India.

Abstract: Artificial intelligence technologies have revolutionized the way we create and manipulate audio, video, images, and text. One of the most notable applications is deepfake content, which uses sophisticated techniques to generate convincing simulations of reality. However, researchers have been developing methods to detect and identify deepfake audio, thus enhancing security in various fields such as media forensics and authentication systems. One such method involves leveraging Mel Spectrograms and Convolutional Neural Networks (CNNs). Mel Spectrograms are visual representations of audio signals that display the frequency components over time. By analyzing these spectrograms, CNNs can be trained to identify patterns and anomalies that indicate artificial alterations in audio content. To develop an effective deepfake detection system, researchers utilized a dataset called Fake-or-Real, which contains a mix of real and deepfake audio samples. The dataset is classified into sub-datasets based on audio length and bit rate, providing a diverse range of samples for comprehensive model training. The trained CNN model can accurately distinguish between real and deepfake audio by identifying subtle or irregularities left behind by deepfake creators. These discrepancies serve as indicators of manipulation and help enhance audio security by automating the detection process. By integrating Mel Spectrograms and CNNs, this approach represents a significant advancement in combating deepfake technology. It offers a promising solution for organizations and individuals looking to protect against misinformation, fraudulent recordings, and other forms of audio manipulation. Moving forward, continued research and refinement of these techniques will further bolster trust and integrity in audio content across various domains, ensuring a safer and more secure digital environment.

Index Terms **–** Deepfake Audio Detection, Mel Spectrogram, Convolutional Neural Network

I. INTRODUCTION

The rise of deepfake technology has posed significant challenges, especially in its manipulation of sound recordings. To address this, developed a comprehensive approach to analyzing deepfake audio. This method involves extracting relevant features from audio recordings, segmenting the data, and labeling it accordingly. Central to this approach is the utilization of Convolutional Neural Networks (CNNs), a type of artificial neural network known for its prowess in analyzing visual data but increasingly being applied to other types of data, such as audio. The CNN plays a crucial role in scrutinizing audio recordings, addressing key challenges like the availability of labeled training data and the computational resources needed for analysis. By harnessing CNNs, this method significantly boosts the accuracy and efficiency of deepfake audio detection. It achieves this by streamlining the detection process, eliminating the need for manual thresholds, and empowering the neural network to autonomously learn and adapt to the complexities of deepfake audio manipulation. This advancement represents a substantial leap forward in combating the proliferation of deceptive audio content enabled by deepfake technology. As technologies in synthetic speech generation continue to advance, audio deepfakes are becoming a prevalent source of deception. Consequently,

distinguishing between fake and real audio is increasingly challenging. The proposed approach relies on optimal feature engineering and the selection of the most effective machine learning models for detecting fake or real audio. Feature engineering encompasses various methods for extracting features from audio, while the feature selection process identifies the minimum set of features that yield the best performance, feeding them into machine learning classifiers. The integration of Mel Spectrograms and CNNs in detecting deepfake audio represents a significant advancement in enhancing audio security, especially in critical areas such as media forensics and authentication systems. Through continued research and refinement, this method holds promise in bolstering trust and integrity in audio content across diverse domains, contributing to a safer and more reliable digital landscape. This research represents a critical step in enhancing audio security and preserving the integrity of audio content in an era where deepfake technology poses significant risks to trust and authenticity.

II. LITERATURE SURVEY

Satpute, Neha Palande, Kishor Jante [1] Deep Fake Voice Detection and Extraction Using Deep Learning a reliable method for identifying deceptive audio content, specifically deepfake voices, using deep learning techniques. It involves three essential steps. Firstly, in the audio preprocessing phase, the input audio data is standardized for consistency in processing, noise reduction techniques are applied to eliminate unwanted background noise, and voice activity detection (VAD) is employed to segment the audio into speech and non-speech regions. Yogesh Patel, Sudeep Tanwar, Pronaya Bhattacharya, Rajesh Gupta, Turki Alsuwian, Innocent Ewean Davidson, Thokozile F. Mazibuko [2] An Improved Dense CNN Architecture for Deepfake Image Detection the aim of this project is to propose an enhanced deep convolutional neural network (D-CNN) architecture for detecting deepfake images with high accuracy and generalizability. A novel D-CNN architecture specifically designed for deepfake image detection. Images from various sources are captured and resized, then fed into the D-CNN model. Binary cross-entropy and Adam optimizer are utilized to enhance the learning rate of the model. The model is trained and tested using seven different datasets containing real and deepfake images from various generative adversarial network (GAN) architectures. Performance metrics such as accuracy and loss values are used for evaluation. Ameer Hamza, Abdul Rehman Javed, Farkhund Iqba, Natalia Kryvinska, Ahmad S. Almadhor, Zunera Jalil, Rouba Borghol [3] Deepfake Audio Detection via MFCC Features Using Machine Learning In this study, the authors use deep and machine learning techniques along with algorithms to detect deepfake audio. For MFCCs technique of audio which provides the most useful information from the MFCCs audio is applied. The experimental results show that the support vector machine (SV atteomized over other machine learning (ML) models for the evaluated accuracy of both datasets "for" and "for2.sec " sec, while the gradiant boosting performed very well for the normalized "for norm" dataset. The VGG-16 model gave very pleasing outcomes when it was tested on the dataset of facials using the original ones. This VGG-16 model exactly will outperform any other current models of the kind. Akash Chintha, Bao Tha, Saniat Javid Sohrawardi, Kartavya Bhatt, Andrea Hickerson, Matthew Wright [4] Recurrent Convolutional Structures for Audio Spoof and Video Deepfake Detection developed effective and efficient forensic methods to detect audio spoof and visual deepfakes using innovative techniques. Developed a robust methodology for the detection of deepfake audiovisual content, which presents significant risks such as defamation of public figures and manipulation of public opinion. Data preprocessing to ensure consistency and remove noise, extract features from audio and visual components using specialized convolutional structures for rich information capture. Leveraging a hybrid architecture comprising bidirectional recurrent structures and entropy-based cost functions, detect spatial and temporal signatures of deepfake renditions. Through rigorous training and evaluation, including comparison with state-of-the-art techniques and extensive generalization studies, aim to advance deepfake detection, providing effective tools to combat deceptive audiovisual content and safeguard against its detrimental impacts on public discourse and perception. Farkhund Iqbal, Ahmed Abbasi, Abdul Rehman Javed, Zunera Jalil and Jamal Al-Karaki [5] Deepfake Audio Detection via Feature Engineering and Machine Learning this paper presents an implementation that is meant to increase the efficiency of modeling using a machine learning classifier for audio files. Firstly, we apply the Principal Component Analysis (PCA) that aims at distilling the most significant features from 270 characteristics associated with each audio file. In summary, this method significantly solves the ill-conditioned problem by taking advantage of computational efficiency as well as achieving the accuracy which results in the performance of audio file analysis.

III. METHODOLOGY

Our proposed methodology included data Collection, pre-processing, segmentation, training, testing and result.

Model Architecture

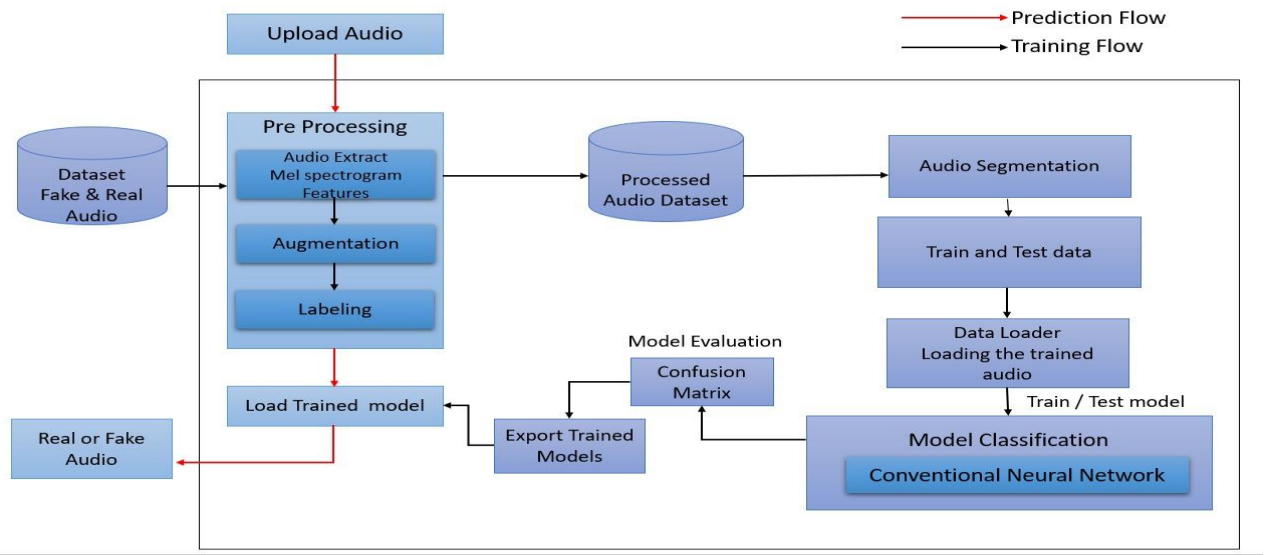

Figure 1: Graphical representation of Proposed approach for deepfake audio detection

Data Collection: Utilized Kaggle's dataset resources for deepfake audio datasets. It contains 2,780 audio files in WAV format. Among these, there are 1,700 fake audio files and 1,080 real audio files, providing a valuable resource for training and testing deepfake audio detection models.

Dataset Link: deep-voice-deepfake-voice-recognition: https://www.kaggle.com/datasets/birdy654/deep-voice-deepfake-voice-recognition

Data Preprocessing: In the preprocessing phase for deepfake audio detection, to enhance the quality and diversity of our dataset while preparing it for training. We begin by extracting Mel spectrograms from raw audio files using library Librosa in Python. Mel spectrograms convert temporal waveforms into two-dimensional representations, capturing frequency content over time crucial for identifying patterns in authentic and fake audio recordings. The extracted audio is shown in figure[2]. Next, we normalize the extracted Mel spectrogram features to ensure consistent scaling across different samples. This normalization improves model convergence during training and prevents biases towards certain amplitude ranges. To increase dataset diversity and robustness, applied data augmentation techniques such as random pitch shifting and time stretching. These augmentations introduce variations, making our model more different acoustic conditions and types of deepfake audio manipulation. By incorporating these preprocessing steps, including Mel spectrogram extraction, data augmentation, and accurate labeling, we prepare our audio data comprehensively. This approach enhances the performance and generalization capabilities of our deepfake audio detection models, ensuring they can accurately identify and distinguish between spoof and bonafide audio recordings.

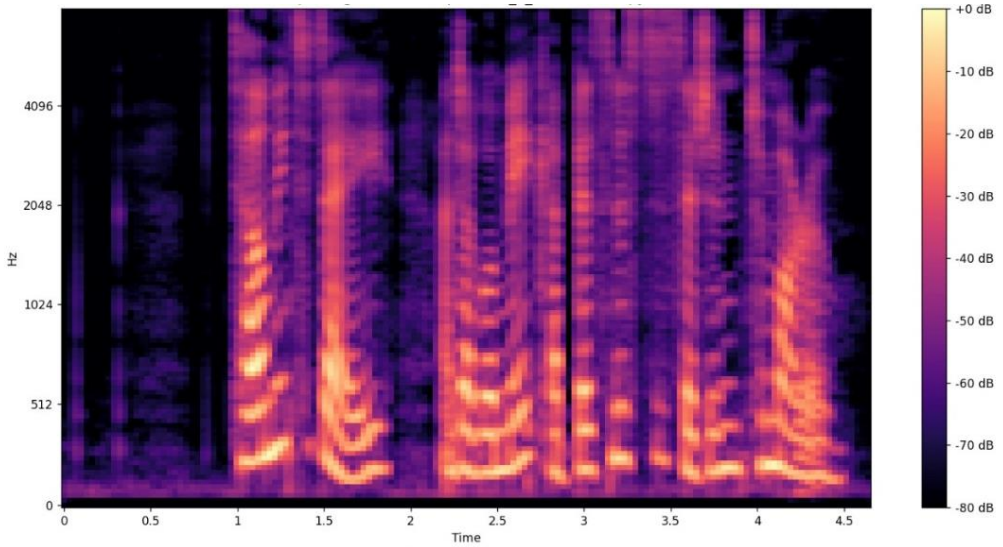

Figure 2: Mel Spectrogram representation of audio signal where the amplitude is depicted in terms of decibel

Segmentation: By breaking down the pre-processed audio into smaller segments, the computational burden is significantly reduce and allowing for more focused and manageable processing, especially with longer recordings or continuous streams. This segmentation technique is complemented by windowing functions applied to each segment by Hamming, which effectively reduce spectral leakage and maintain precise frequency representation within segments. Overlapping frames between adjacent segments are crucial for maintaining continuity and capturing temporal dependencies across segments. This ensures smoother transitions for robust feature extraction. However, the percentage of overlap must be carefully determined to strike a balance between computational efficiency and temporal continuity. Higher overlap percentages offer smoother transitions but increase computational complexity, while lower percentages reduce computational overhead but may compromise continuity. Audio segmentation optimizes the processing of deepfake audio by breaking it into manageable chunks, employing windowing to minimize spectral leakage, utilizing overlapping frames for continuity, and balancing overlap percentages for efficient analysis. These techniques collectively contribute to more accurate feature extraction and robust detection of deepfake audio content.

.

CLASSIFICATION MODEL

CONVOLUTIONAL NEURAL NETWORK (CNN): The CNN model employed for audio classification leverages the power of convolutional layers to process spectrogram images, treating them as image data. This approach allows the model to utilize image-based techniques for analyzing the frequency content of audio signals represented in spectrograms.

The formula for a typical convolutional layer in a CNN is as follows:

$Z[l]=W[l]*A[l-1]+b[l]$

Where: $Z[l]$ is the output of the convolutional layer, $W[l]$ represents the weights associated with the convolutional filters, $A[l-1]$ denotes the input to the convolutional layer, $b[l]$ is the bias term.

By utilizing convolutional layers, the CNN model excels at capturing intricate patterns within spectrogram data through local receptive field operations. This means that the model focuses on small, localized areas of the spectrogram at a time, allowing it to detect detailed features that are crucial for audio classification tasks. During training, the CNN model learns to differentiate between bonafide (genuine) and spoof (fake) audio signals by processing a labeled dataset. This dataset contains spectrogram images corresponding to both bonafide and spoof audio signals, allowing the model to learn the characteristics that distinguish these two classes. The CNN model extracts hierarchical features from spectrogram data, encompassing both low-level details and high-level representations. This hierarchical feature extraction is achieved through multiple convolutional layers, where each layer extracts progressively abstract features from the spectrogram images.

The Softmax activation function is commonly used in CNNs for multi-class classification tasks. It transforms the output of the last convolutional layer into a probability distribution over the different classes.

The Softmax function exponentiates the output values and then normalizes them by dividing by the sum of all exponentiated outputs. This normalization ensures that the output values lie between 0 and 1 and sum up to 1, effectively representing probabilities for each class. The class with the highest probability is then chosen as the predicted class by the CNN model during inference.

The CNN model's formula and architecture enable it to effectively analyze spectrogram data, extract relevant features, and discern complex patterns indicative of bonafide or spoof audio signals, contributing to accurate audio classification in deepfake detection systems.

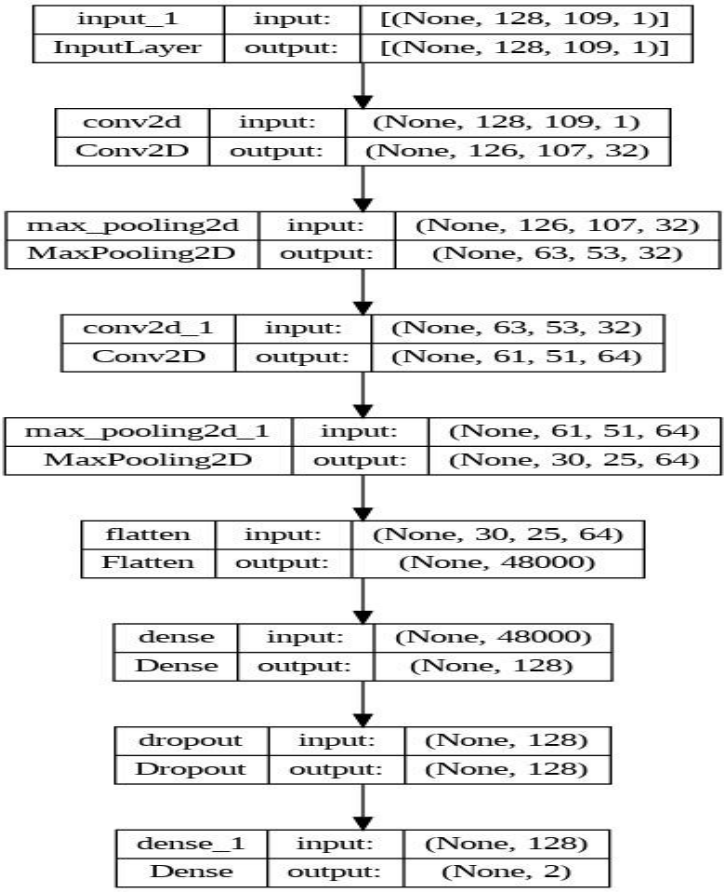

Figure 3: CNN Model Architecture

In figure 3 the CNN architecture is designed with two convolutional layers followed by max-pooling layers to extract features from the input spectrogram data. The Flatten layer is used to transform the output from convolutional layers into a flat vector, which is then passed through a dense layer with a ReLU activation function to learn higher-level features. Finally, the output layer uses a softmax activation function for multi-class classification, outputting probabilities for each class. The model is compiled using the Adam optimizer, categorical cross-entropy loss function (suitable for multi-class classification), and accuracy as the evaluation metric. This compilation step prepares the model for training with the specified hyperparameters and optimization settings.

EXPERIMENT

Dataset Preparation: The dataset used consists of audio samples in FLAC format. The labels for each audio file are obtained from the file "ASVspoof2019.LA.cm.train.trn.txt". The label indicates whether the audio is bonafide or fake. Audio files are loaded using librosa and converted into Mel Spectrograms to extract relevant features for training the CNN model.

Model Training: The CNN model architecture is defined with input shape (128, 109, 1) corresponding to the Mel Spectrogram dimensions. The model consists of Conv2D layers with ReLU activation, MaxPooling2D layers, a Flatten layer, and Dense layers with dropout for regularization. The categorical cross-entropy loss function and Adam optimizer are used to compile the model.

Training Procedure: The training data is split into batches of size 32 and trained for 10 epochs. During training, the model learns to distinguish between real and fake audio samples based on the features extracted from Mel Spectrograms. The model's performance is evaluated using accuracy metrics on both the training and validation sets.

Model Evaluation: After training, the model is saved as "audio_classifier1.h5" for later use. The trained model is then loaded in the "test.py" script to perform inference on unseen audio data. Saving the trained model is crucial as it allows for later use without the need to retrain the model from scratch every time it's needed.

Inference and Detection: An audio file is selected for testing in the GUI application. The selected audio file is loaded, and its Mel Spectrogram is computed using librosa. The Mel Spectrogram is passed through the

trained CNN model for classification as either real or fake. The model's prediction, confidence level, and accuracy metrics (real/fake) are displayed in the GUI interface. Additionally, a graphical representation of the Mel Spectrogram is shown to visualize the audio data's features.

The experiment demonstrates the effectiveness of using CNNs and Mel Spectrograms for deepfake audio detection. The model shows promising results in accurately classifying unseen audio samples and provides confidence levels and accuracy metrics to assess the classification reliability. Overall, the experiment contributes to the advancement of deepfake detection techniques in audio data.

Evaluation Process: The model's performance is evaluated on a separate test dataset by comparing its predictions with ground truth labels. Accuracy is calculated as the ratio of correctly classified audio samples to the total number of samples in the test dataset. Additionally, the confidence level associated with each prediction is analyzed to gain insights into the model's performance.

Accuracy Calculation: Accuracy is determined by comparing the model's predictions with the ground truth labels in the test dataset. The accuracy calculation formula is: Accuracy = (Number of Correct Predictions / Total Number of Predictions) × 100%.

Confidence Level Calculation: The confidence level for a prediction is derived from the output probabilities assigned to each class (real or fake). It is calculated as the maximum of the probabilities assigned to the predicted classes. A higher confidence level indicates greater certainty in the model's prediction.

Classification: After passing segmented frames through the CNN layers, the output is flattened and passed through fully connected layers. A softmax activation function is used in the output layer to obtain probabilities for each class (real or fake).During training, appropriate loss functions (e.g., categorical cross-entropy) and optimization algorithms (e.g., Adam) are applied to update the model's parameters.

IV. EXISTING SYSTEM

The existing system described integrates deep learning models with traditional handcrafted features to address the complexity of detecting deepfake audio. The hybrid architecture of VGG16 and LSTM combines the strengths of a CNN renowned for image classification and an RNN designed for sequential data processing. This integration ensures sensitivity to both spatial cues and temporal dynamics in the audio data, enhancing the ability to distinguish between genuine and fabricated recordings.

A comprehensive set of features is generated from the sound signals using the deep learning models and a feature ensembling strategy, which includes MFCC-40 coefficients and various acoustic characteristics like roll-off point and centroid. By integrating and exploring various feature extraction techniques, the system provides a holistic representation of the audio data, thereby improving classification precision.

Machine learning methods such as SVM, RF, KNN, and XGB are employed for classification across different fake and real dataset noise environments and various audio scenarios. The system demonstrates high resistance to the introduction of deepfake audio into other contexts, achieving impressive accuracies of 83% with the VGG16 model and 89% with the LSTM model.

V. PROPOSED SYSTEM

The proposed system for detecting fake audio integrates deep learning algorithm and advanced signal processing techniques to address the growing threat posed by sophisticated audio manipulation tools. With the emergence of technologies like "deepfake audio" and "voice cloning". To combat these challenges and reinforce trust in audio-based applications, our system employs a multi-faceted approach:

Combination of Machine Learning Algorithms and Signal Processing Techniques:

Machine learning algorithms offer the ability to learn patterns and features from data, which can be invaluable in distinguishing between authentic and manipulated audio recordings. By leveraging these algorithms, your system can automatically detect anomalies or discrepancies indicative of manipulation.

Signal processing techniques, on the other hand, provide tools for analyzing the underlying characteristics of audio signals. Techniques like mel spectrograms and segmentation windowing with overlapping enable the extraction of relevant features from audio data, which can aid in identifying subtle differences between real and fake audio.

Addressing the Challenge of Deepfake Audio:

Deepfake audio and voice cloning technologies pose a significant threat due to their ability to produce convincing, lifelike audio recordings of individuals saying or doing things they never actually did. To counter this, your system employs a combination of techniques aimed at detecting the anomalies characteristic of deepfake audio.

By analyzing features such as spectral patterns, temporal characteristics, and consistency of voice characteristics, your system can identify discrepancies that indicate potential manipulation.

Mel Spectrograms and Segmentation Windowing:

Mel spectrograms are a representation of the spectrum of a signal as it varies over time. They are particularly useful in audio signal processing tasks as they highlight the frequency components of the audio signal.

Segmentation windowing with overlapping involves dividing the audio signal into smaller segments and applying a window function to each segment. Overlapping segments help in capturing temporal information and smoothing out transitions between segments, which can improve the accuracy of feature extraction.

Integration of Convolutional Neural Networks (CNNs):

CNNs are a type of deep learning algorithm known for their effectiveness in analyzing spatial data such as images. However, they can also be applied to sequential data like audio signals.

By leveraging CNNs, your system can automatically learn hierarchical features directly from raw audio input data. This ability to extract intricate patterns and features from audio signals enhances the system's capability to discern between authentic and manipulated audio recordings.

Enhanced Detection of Fake Audio:

The comprehensive approach taken by your system, combining machine learning algorithms, signal processing techniques, and CNNs, strengthens the defense against the proliferation of sophisticated audio manipulation tools.

By accurately identifying manipulated or synthetic audio recordings, your system reinforces trust and dependability in audio-based applications, thereby mitigating the risks associated with fake audio recordings such as fraud, reputation damage, and the dissemination of misinformation.

VI. RESULT

The trained model shows the most accurate result in the deepfake audio detection with audio accuracy: 0.9508464693536824, overall confidence: 0.9649833786365379

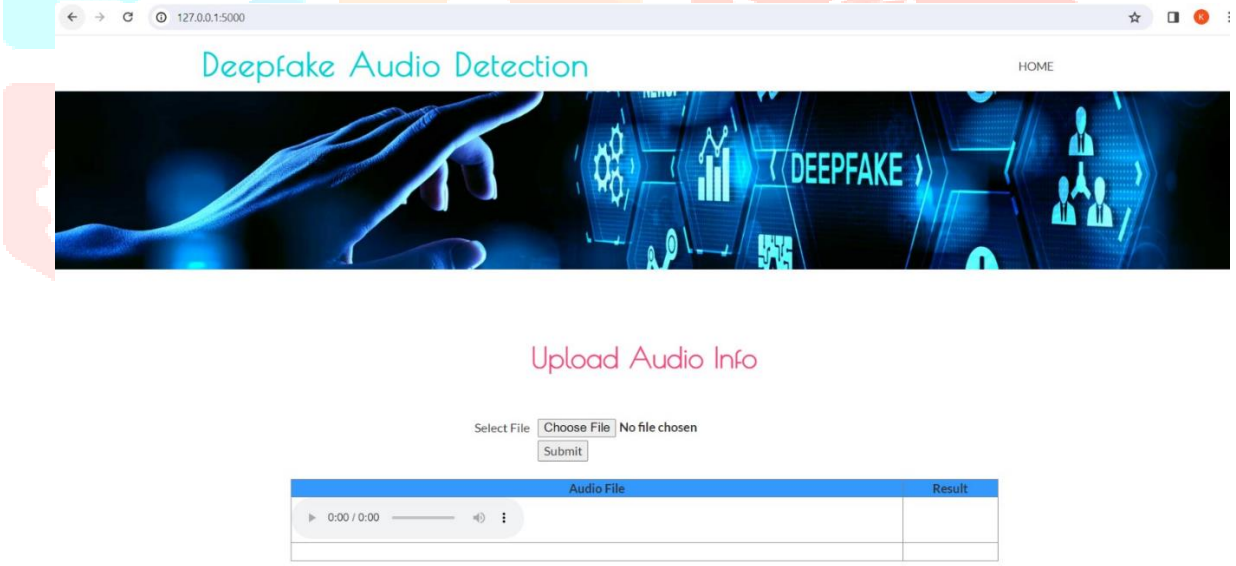

Figure 4: Audio Upload Page

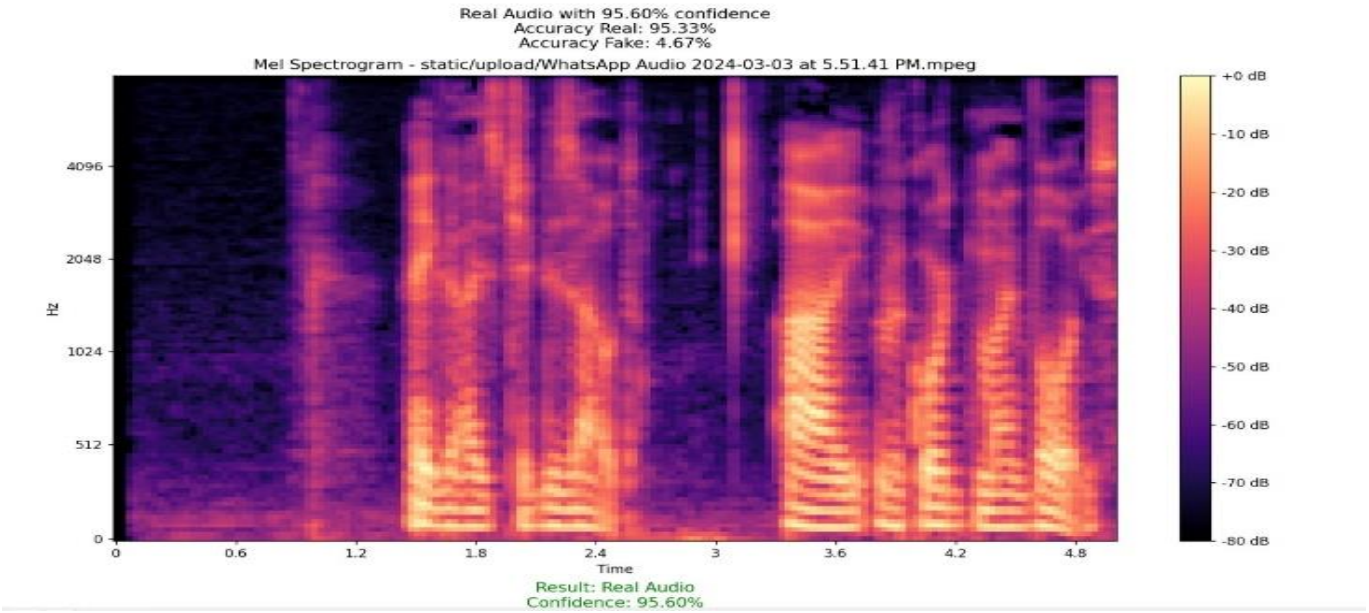

Figure 5: Bonafide audio result

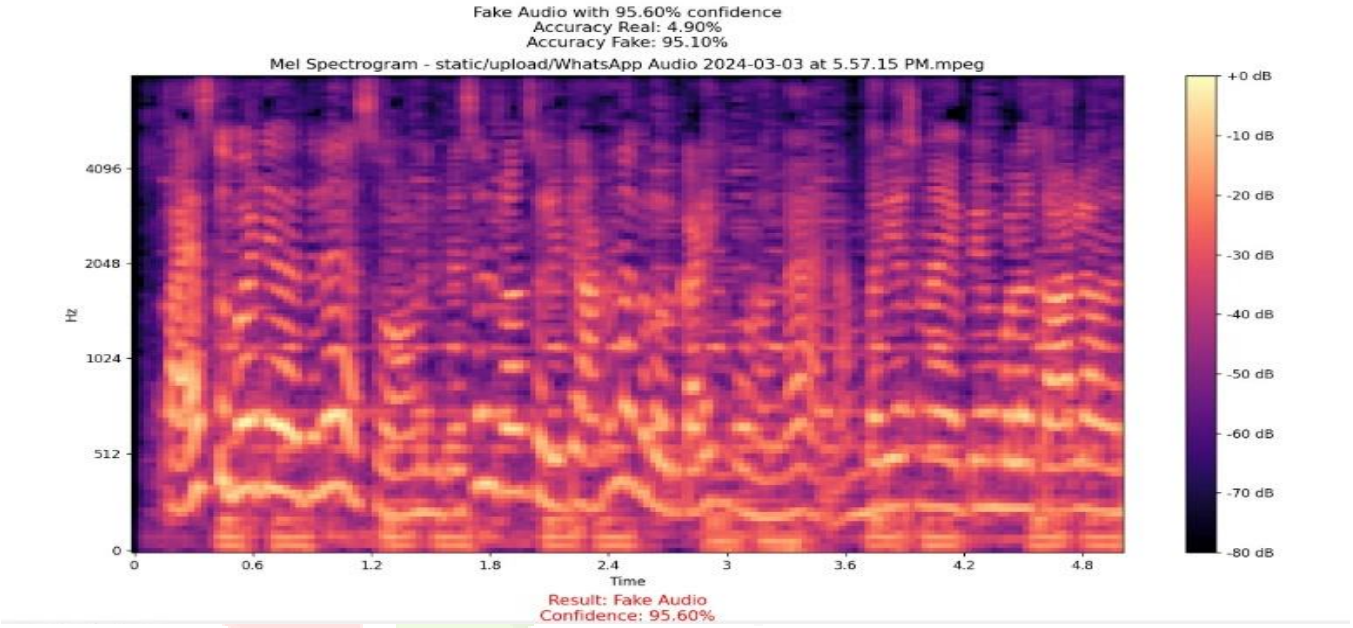

Figure 6: Spoof audio result

VII. CONCLUSION

In conclusion, the identification of deepfake audio poses a significant challenge within the domain of digital forensics and media integrity verification. With the continuous advancement of artificial intelligence and machine learning, deepfake technologies are becoming increasingly sophisticated and widespread. However, amidst these complexities, ongoing research and development efforts are focused on devising effective strategies for detecting and countering the dissemination of altered audio content. Various techniques are being explored, including the analysis of spectral patterns, identification of inconsistencies in speech characteristics, and the utilization of blockchain technology for immutable verification. These approaches show promise in addressing the issue by enhancing the ability to identify manipulated audio materials. It's crucial to recognize that while these detection methods provide a level of defense against the proliferation of deepfake audio, they are not infallible and require continuous enhancement to keep up with evolving manipulation tactics. Moreover, the ethical considerations surrounding the use of deepfake technology highlight the importance of promoting media literacy and responsible sharing of information. Ongoing efforts in research and technological innovation are essential to combat the challenges posed by deepfake audio, emphasizing the need for vigilance, adaptability, and ethical awareness in navigating this complex landscape.

VIII. REFERENCES

[1] Waseem, Saima, Syed R. Abu-Bakar, Bilal Ashfaq Ahmed, Zaid Omar, Taiseer Abdalla Elfadil Eisa, and Mhassen Elnour Elneel Dalam. "DeepFake on Face and Expression Swap: A Review." IEEE Access (2023).
[2] Abbasi, Ahmed, Abdul Rehman Rehman Javed, Amanullah Yasin, Zunera Jalil, Natalia Kryvinska, and Usman Tariq. "A large-scale benchmark dataset for anomaly detection and rare event classification for audio forensics." IEEE Access 10 (2022): 38885-38894.
[3] V. Phani and Krishna Deep, "Fake detection using LSTM and RESNEXT", *Journal of Engineering Sciences*, vol. 13, no. 07, July 2022, ISSN 0377–9254.
[4] Pramod Dhamdhere, "Semantic trademark retrieval system based on conceptual similarity of text with leveraging histogram computation for images to reduce trademark infringement", Webology (ISSN: 1735-188X), Volume 18, Number 5, 2021.
[5] J. Khochare, C. Joshi, B. Yenarkar, S. Suratkar, F. Kazi, A deep learning framework for audio deepfake detection, Arabian Journal for Science and Engineering (2021) 1–12.
[6] D. Cozzolino, M. Nießner and L. Verdoliva, "Audio-visual person-of-interest deepfake detection", 2022.
[7] angyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, et al., "Add 2022: the first audio deep synthesis detection challenge", *2022 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2022.
[8] Y. Gao, T. Vuong, M. Elyasi, G. Bharaj and R. Singh, "Generalized Spoofing Detection Inspired from Audio Generation Artifacts", 2021.
[9] R. Yamamoto, E. Song, and J. Kim. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In ICASSP, 2020.
[10] K. Chugh, P. Gupta, A. Dhall and R. Subramanian, "Not made for each other-audio-visual dissonance-based deepfake detection and localization", *arXiv: Computer Vision and Pattern Recognition*, 2020.
[12] E. AlBadawy, S. Lyu and H. Farid, "Detecting ai-synthesized speech using bispectral analysis", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.
[13] E. AlBadawy, S. Lyu and H. Farid, "Detecting ai-synthesized speech using bispectral analysis", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.
[14] W. Cai, J. Chen and M. Li, "Exploring the encoding layer and loss function in end-to-end speaker and language recognition system", 2018.