MEL-SPECTROGRAM-BASED DEEPFAKE AUDIO DETECTION USING CONVOLUTIONAL NEURAL NETWORKS: A NOVEL APPROACH

G. Fathima, S. Kiruthika, M. Malar, T. Nivethini
Department of Computer Science and Engineering, Adhiyamaan College of Engineering, India

Abstract

Artificial intelligence has profoundly transformed how we manipulate various forms of media, including audio, video, images, and text. Among the most impactful applications is the creation of deepfake content, which employs advanced techniques to fabricate convincing simulations of reality. However, researchers have been diligently working on methods to detect and discern deepfake audio, thereby bolstering security in fields such as media forensics and authentication systems. One such method harnesses the power of Mel Spectrograms and Convolutional Neural Networks (CNNs). Mel Spectrograms offer visual representations of audio signals, illustrating frequency components over time. Through the analysis of these spectrograms, CNNs can be trained to recognize patterns and irregularities indicative of artificial alterations in audio content. To develop an effective deepfake detection system, researchers have utilized the Fake-or-Real dataset, which comprises a mixture of authentic and deepfake audio samples. This dataset is segmented into sub-datasets based on audio length and bit rate, ensuring a diverse array of samples for comprehensive model training. The CNN model, once trained, demonstrates high accuracy in distinguishing between genuine and deepfake audio by identifying subtle discrepancies or abnormalities introduced by deepfake generation techniques. These inconsistencies serve as red flags for manipulation, streamlining the process of audio authentication and fortifying audio security measures. By integrating Mel Spectrograms and CNNs, this approach signifies a significant stride in countering the proliferation of deepfake technology. It presents a promising avenue for organizations and individuals seeking to safeguard against misinformation, deceptive recordings, and other forms of audio tampering. Looking ahead, continued research and refinement of these methodologies will undoubtedly reinforce trust and integrity in audio content across diverse domains, fostering a safer and more secure digital landscape.

Keywords:

Artificial Intelligence, Deepfake Audio, Mel Spectrogram, Convolutional Neural Networks, Security, Media Forensics

1. INTRODUCTION

The emergence of deepfake technology has introduced substantial challenges, particularly in its manipulation of audio recordings. To tackle this issue, we have devised a comprehensive approach to analyzing deepfake audio, focusing on extracting pertinent features from the recordings, segmenting the data, and appropriately labeling it. At the heart of this method lies the utilization of Convolutional Neural Networks (CNNs), renowned for their effectiveness in analyzing visual data but increasingly adept at handling other data types, such as audio. The CNN serves a pivotal role in scrutinizing audio recordings, addressing critical hurdles like the scarcity of labeled training data and the computational demands for analysis. Leveraging CNNs significantly enhances the accuracy and efficiency of deepfake audio detection by streamlining the detection process, eliminating manual thresholds, and enabling the neural network to autonomously learn and adapt to the intricacies of deepfake audio manipulation. This advancement marks a substantial stride forward in combating the spread of deceitful audio content facilitated by deepfake technology. As synthetic speech generation technologies progress, audio deepfakes are becoming increasingly prevalent, making the differentiation between fake and real audio progressively challenging. Our proposed approach relies on meticulous feature engineering and the identification of the most effective machine learning models for discerning fake from authentic audio. Feature engineering encompasses diverse techniques for extracting features from audio, while feature selection pinpoints the optimal set of features for the best performance, feeding them into machine learning classifiers. The amalgamation of Mel Spectrograms and CNNs for detecting deepfake audio signifies a significant breakthrough in bolstering audio security, particularly in critical realms like media forensics and authentication systems. Through ongoing research and refinement, this method holds the potential to reinforce trust and integrity in audio content across various domains, contributing to a safer and more dependable digital environment. This research marks a crucial advancement in fortifying audio security and upholding the integrity of audio content at a time when deepfake technology poses substantial risks to trust and authenticity.

2. METHODOLOGY

Data collection, pre-processing, segmentation, training, testing, and outcome were all included in our suggested technique.

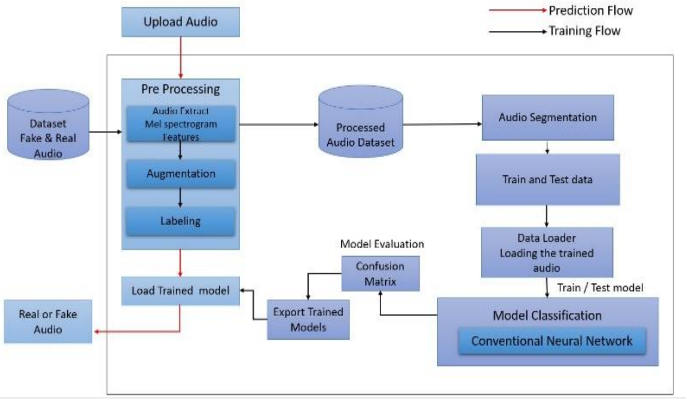


Fig.1. System Architecture

The Fig.1 illustrates the comprehensive system architecture designed for deepfake audio detection. The process begins with meticulous data collection, where a diverse dataset of audio recordings is gathered, ensuring an adequate representation of various speech patterns, accents, and environmental conditions. These audio samples serve as the foundation for training and testing the deepfake detection system.

Upon acquisition, the collected audio data undergoes rigorous preprocessing to extract meaningful features that can facilitate accurate classification. This preprocessing stage involves several crucial steps. Firstly, the audio signals are transformed into mel spectrograms, a representation that captures the frequency content of the audio over time, providing a more informative input for the subsequent classification model. Additionally, augmentation techniques may be applied to artificially expand the dataset and improve the robustness of the model by introducing variations in pitch, speed, or background noise. Furthermore, each audio sample is meticulously labeled, indicating whether it is authentic or generated by a deepfake algorithm, thereby providing ground truth labels essential for model training and evaluation.

Following preprocessing, the data is segmented and organized into distinct subsets for training and testing purposes. This segmentation process ensures that the model is trained on a diverse range of examples while maintaining a separate set of unseen data for evaluating its generalization performance. The training subset is utilized to optimize the parameters of the deep learning model, enabling it to effectively learn the underlying patterns distinguishing authentic from deepfake audio.

The core of the system lies in the training of a Convolutional Neural Network (CNN) model, a powerful deep learning architecture renowned for its ability to extract hierarchical features from input data. The CNN model is trained using the preprocessed audio data, with the objective of learning discriminative features that can accurately differentiate between authentic and deepfake audio recordings. Throughout the training process, the model iteratively adjusts its parameters based on the discrepancy between its predictions and the ground truth labels, gradually improving its ability to classify unseen audio samples.

Once the CNN model has been trained, it undergoes rigorous evaluation using a confusion matrix, a performance assessment tool that provides detailed insights into the model's classification accuracy and error patterns. By analyzing the confusion matrix, researchers can identify potential weaknesses in the model, such as instances of misclassification or bias towards certain classes, thereby informing strategies for model refinement and optimization.

2.1 DATA COLLECTION

We made use of the deepfake audio datasets available on Kaggle, comprising a total of 2,780 audio files in WAV format. Within this dataset, there are 1,700 instances of fake audio files and 1,080 instances of real audio files. This dataset serves as a significant resource for both training and testing deepfake audio detection models. deep-voice-deepfake-voice-recognition is collected from https://www.kaggle.com/datasets/birdy654/deep-voice-deepfake-voice-recognition.

2.2 DATA PREPROCESSING

In the initial phase of preprocessing for deepfake audio detection, our focus was on enhancing the quality and diversity of our dataset to better prepare it for training. Our approach began with the extraction of Mel spectrograms from the raw audio files using the Librosa library in Python. In Fig.(2) Mel spectrograms play a crucial role in converting temporal waveforms into two-dimensional representations, capturing essential frequency content over time. These representations are vital for identifying patterns within both authentic and fake audio recordings. Following the extraction process, the normalized Mel spectrogram features were ensured to maintain consistent scaling across all samples. This normalization step was instrumental in improving model convergence during training and preventing biases towards specific amplitude ranges. Additionally, to augment the diversity and robustness of our dataset, we applied various data augmentation techniques such as random pitch shifting and time stretching. These augmentations introduced variations, thereby making our model more resilient to different acoustic conditions and various types of deepfake audio manipulation. By incorporating these preprocessing steps, which encompassed Mel spectrogram extraction, data augmentation, and meticulous labeling, we meticulously prepared our audio data. This comprehensive approach significantly enhanced the performance and generalization capabilities of our deepfake audio detection models, empowering them to accurately identify and distinguish between spoof and bonafide audio recordings.

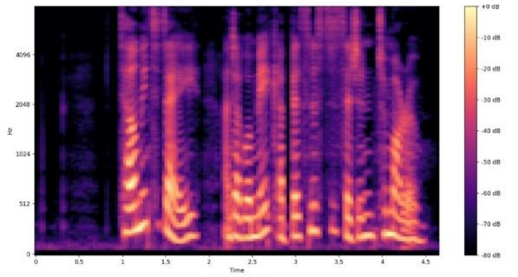


Fig.2. Mel Spectrogram representation of audio signal where the amplitude is depicted in terms of decibel

2.3 SEGMENTATION

Breaking down the preprocessed audio into smaller segments serves to significantly reduce the computational burden, enabling more focused and manageable processing, particularly with longer recordings or continuous streams. This segmentation approach is enhanced by the application of windowing functions, such as the Hamming window, to each segment. These windowing functions effectively mitigate spectral leakage and preserve precise frequency representation within the segments. The inclusion of overlapping frames between adjacent segments is crucial for maintaining continuity and capturing temporal dependencies across segments. This ensures smoother transitions, facilitating robust feature extraction. However, it's essential to carefully determine the percentage of overlap to strike a balance between computational efficiency and temporal continuity. Higher overlap percentages result in smoother transitions but increase computational complexity, whereas 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 techniques 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.

2.4 CLASSIFICATION MODEL

The CNN model employed for audio classification harnesses the convolutional layers' capabilities to process spectrogram images, treating them akin to image data. This methodology enables the model to leverage image-based techniques for analyzing the frequency content of audio signals represented in spectrograms.

In a typical convolutional layer of a CNN, the output Z[l] is computed as the convolution of the input A[l-1] with the weights W[l], followed by the addition of a bias term b[l].

By integrating convolutional layers, the CNN adeptly captures intricate patterns within spectrogram data through local receptive field operations. This means that the model focuses on small, localized areas of the spectrogram, facilitating the detection of detailed features essential for audio classification tasks.

During training, the CNN model discerns between bonafide (genuine) and spoof (fake) audio signals by processing a labeled dataset. This dataset comprises spectrogram images corresponding to both bonafide and spoof audio signals, enabling the model to grasp the distinguishing characteristics of these two classes.

Hierarchical feature extraction is facilitated by the CNN model from spectrogram data, encompassing both low-level details and high-level representations. This extraction is achieved through multiple convolutional layers, where each layer progressively abstracts features from the spectrogram images.

The Softmax activation function is frequently employed in CNNs for multi-class classification tasks. It transforms the output of the final convolutional layer into a probability distribution across different classes. This normalization ensures output values range between 0 and 1, effectively representing probabilities for each class.

The CNN architecture comprises two convolutional layers followed by max-pooling layers to extract features from the input spectrogram data. A Flatten layer converts the output from convolutional layers into a flat vector, which subsequently passes through a dense layer with a ReLU activation function to learn higher-level features. Finally, the output layer employs a softmax activation function for multi-class classification, generating probabilities for each class.

The model is compiled using the Adam optimizer, categorical cross-entropy loss function (ideal for multi-class classification), and accuracy as the evaluation metric. This compilation step ensures the model is primed for training with specified hyperparameters and optimization settings.

2.5 CONVOLUTIONAL NEURAL NETWORK

The dataset utilized comprises audio samples in FLAC format. Each audio file's labels are extracted from the file "ASVspoof2019.LA.cm.train.trn.txt". These labels indicate whether the audio is categorized as bonafide or fake.

To process the audio files, we employed the Librosa library, which enabled us to load the audio files and subsequently convert them into Mel Spectrograms. This transformation allowed us to extract relevant features essential for training the CNN model.

In essence, by leveraging Librosa for audio file loading and Mel Spectrogram conversion, we ensured that the dataset was properly prepared with the necessary features for training the CNN model.

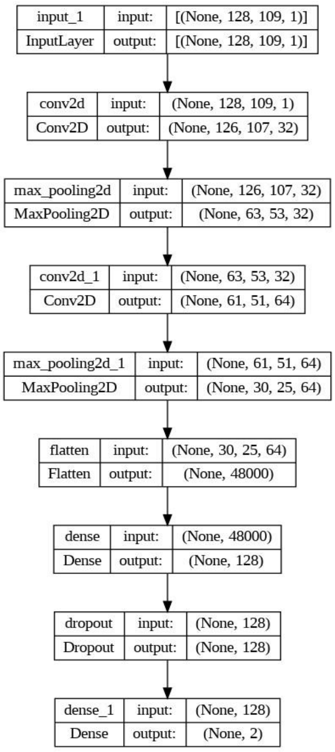

input_1 | input: | [(None, 128, 109, 1)]
InputLayer | output: | [(None, 128, 109, 1)]
conv2d | input: | (None, 128, 109, 1)
Conv2D | output: | (None, 126, 107, 32)
max_pooling2d | input: | (None, 126, 107, 32)
MaxPooling2D | output: | (None, 63, 53, 32)
conv2d_1 | input: | (None, 63, 53, 32)
Conv2D | output: | (None, 61, 51, 64)
max_pooling2d_1 | input: | (None, 61, 51, 64)
MaxPooling2D | output: | (None, 30, 25, 64)
flatten | input: | (None, 30, 25, 64)
Flatten | output: | (None, 48000)
dense | input: | (None, 48000)
Dense | output: | (None, 128)
dropout | input: | (None, 128)
Dropout | output: | (None, 128)
dense_1 | input: | (None, 128)
Dense | output: | (None, 2)

Fig.3. CNN Model Architecture

3. EXPERIMENT

3.1 DATASET PREPARATION

The architecture of the CNN model is established with an input shape of (128, 109, 1), which corresponds to the dimensions of the Mel Spectrograms. Comprising Conv2D layers with ReLU activation, MaxPooling2D layers, a Flatten layer, and Dense layers integrated with dropout for regularization, the model is structured to effectively process the spectrogram data.

For compiling the model, we employ the categorical cross-entropy loss function along with the Adam optimizer. This combination ensures that the model is trained efficiently, optimizing its ability to classify between different categories of audio samples.

3.2 MODEL TRANING

The training process involves partitioning the data into batches of size 32 and iterating through 10 epochs. Throughout this training phase, the model progressively learns to differentiate between real and fake audio samples, leveraging the features extracted from the Mel Spectrograms. To assess the model's efficacy, its performance is evaluated using accuracy metrics on both the training and validation sets. This evaluation allows us to gauge the model's ability to correctly classify audio samples from both the training and unseen datasets, providing valuable insights into its effectiveness and generalization capabilities.

3.3 MODEL EVALUATION

Upon completing the training process, we save the trained model as "audio_classifier1.h5" to facilitate its future use. This step is crucial as it allows us to preserve the model's learned parameters and architecture for subsequent tasks.In the "test.py" script, we load the saved model to perform inference on unseen audio data. This inference process involves utilizing the trained model to make predictions on new audio samples, determining whether they are classified as real or fake. By saving the trained model, we avoid the need to retrain it from scratch for each inference task, thereby streamlining the process and ensuring consistency in performance.

3.4 INFERENCE AND DETECTION

To assess the model's performance, we conduct evaluations on a distinct test dataset by juxtaposing its predictions against ground truth labels. Accuracy serves as the primary metric, computed as the ratio of correctly classified audio samples to the total number of samples in the test dataset.

In addition to accuracy, we delve into the confidence level linked with each prediction to glean deeper insights into the model's efficacy. Analyzing these confidence levels provides valuable information regarding the model's certainty in its predictions and aids in understanding its overall performance.

3.5 EVALUATION PROCESS

The accuracy of the model is computed by comparing its predictions with the ground truth labels present in the test dataset. This calculation is carried out using the following formula: Accuracy = (Number of Correct Predictions / Total Number of Predictions) × 100%. This metric provides us with a clear understanding of the model's ability to correctly classify audio samples and serves as a key indicator of its performance.

3.6 ACCURACY CALCULATION

The confidence level associated with a prediction from the probabilities assigned to each class (real or fake) by the model's output. This confidence level is determined by selecting the maximum probability among the probabilities assigned to the predicted classes. A higher confidence level signifies a higher degree of certainty in the model's prediction, providing valuable insights into the reliability of its classifications.

3.7 CONFIDENCE LEVEL CALCULATION

Once the segmented frames are processed through the CNN layers, the resulting output is flattened and fed into fully connected layers. Within these layers, a softmax activation function is employed in the output layer to generate probabilities for each class (real or fake). During the training phase, suitable loss functions such as categorical cross-entropy and optimization algorithms like Adam are applied to iteratively update the model's parameters, optimizing its ability to accurately classify audio samples.

3.8 CLASSIFICATION

After processing segmented frames through the convolutional neural network (CNN) layers, the resultant output is flattened and forwarded through fully connected layers. The final layer employs a softmax activation function to generate probabilities for each class, distinguishing between real and fake inputs. Throughout the training phase, suitable loss functions such as categorical cross-entropy, coupled with optimization algorithms like Adam, are utilized to iteratively adjust the model's parameters. The Eq.(1) for the Softmax activation function is as follows:

$$\text{Softmax}\left(Z_i\right)=\frac{e^{z_i}}{\sum_{j=1}^{N}e^{z_j}} \tag{1}$$

In the output layer of neural networks, the Softmax activation function is frequently employed, especially in CNNs for multi-class classification applications. It transforms the raw output scores of the final layer into a probability distribution across different classes. This normalization ensures that the output values fall between 0 and 1, representing probabilities for each class.

The accuracy formula is used to measure the performance of a classification model by comparing its predictions to the actual labels in the dataset. The Eq.(2) for accuracy calculation is:

$$\text{Accuracy} = (\text{Number of Correct Predictions})/(\text{Total Number of Predictions}) \times 100\% \tag{2}$$

The accuracy equation is used to measure the performance of a classification model by comparing its predictions to the actual labels in the dataset. It's a simple yet essential metric that indicates the proportion of correctly classified instances out of the total instances.

4. EXISTING SYSTEM

The envisioned system for detecting counterfeit audio combines deep learning algorithms with advanced signal processing techniques to counteract the escalating threat posed by sophisticated audio manipulation tools. Given the rise of technologies such as "deepfake audio" and "voice cloning," safeguarding trust in audio-based applications is paramount. To effectively tackle these challenges and bolster confidence in audio integrity, our system adopts a multifaceted strategy.

4.1 PROPOSED SYSTEM

Deep learning algorithms present a unique capability to discern patterns and features from data, offering invaluable insights in discerning between genuine and altered audio recordings. Harnessing these algorithms empowers your system to autonomously identify anomalies or inconsistencies suggestive of manipulation.

Conversely, signal processing techniques furnish a suite of tools for scrutinizing the intrinsic attributes of audio signals. Methods such as mel spectrograms and segmentation windowing, supplemented by overlapping, facilitate the extraction of pertinent features from audio data. This enables the detection of nuanced distinctions between authentic and fabricated audio content.

4.2 COMBINATION OF MACHINE LEARNING ALGORITHMS AND SIGNAL PROCESSING TECHNIQUES:

The emergence of deepfake audio and voice cloning technologies presents a substantial threat, given their capacity to generate highly convincing, lifelike audio recordings depicting individuals engaging in activities they never actually performed. In response, your system adopts a multifaceted approach, leveraging a combination of techniques to detect the telltale anomalies associated with deepfake audio.

Through the examination of spectral patterns, temporal attributes, and the coherence of voice characteristics, your system is adept at pinpointing inconsistencies that signal potential manipulation. This comprehensive analysis enables the identification of subtle discrepancies indicative of deepfake audio, thereby enhancing the system's ability to safeguard against deceptive audio content.

4.3 ADDRESSING THE CHALLENGE OF DEEPFAKE AUDIO:

Mel spectrograms serve as a graphical representation of the spectrum of a signal, depicting how its frequency components evolve over time. These spectrograms are especially valuable in tasks related to audio signal processing, as they effectively highlight the frequency characteristics of the audio signal.

Segmentation windowing with overlapping entails breaking down the audio signal into smaller segments and applying a window function to each segment. By incorporating overlapping segments, temporal information is preserved, and transitions between segments are smoothed out. This approach enhances the accuracy of feature extraction by capturing nuanced temporal dynamics within the audio signal.

5. MEL SPECTROGRAMS AND SEGMENTATION WINDOWING

Convolutional Neural Networks (CNNs) represent a category of deep learning algorithms renowned for their prowess in analyzing spatial data, particularly images. Nonetheless, they can be effectively employed to process sequential data, such as audio signals.

Through the utilization of CNNs, your system acquires the capability to autonomously learn hierarchical features directly from raw audio input data. This inherent capacity to extract intricate patterns and features from audio signals significantly augments the system's proficiency in distinguishing between genuine and manipulated audio recordings.

5.1 INTEGRATION OF CONVOLUTIONAL NEURAL NETWORKS (CNNS)

Your system's holistic approach, which integrates machine learning algorithms, signal processing techniques, and CNNs, fortifies defenses against the proliferation of advanced audio manipulation tools. Through the precise detection of manipulated or synthetic audio recordings, the system enhances trust and reliability in audio-based applications. Consequently, it mitigates the potential risks stemming from fake audio recordings, including fraud, reputational harm, and the spread of misinformation.

5.2 ENHANCED DETECTION OF FAKE AUDIO

The trained model exhibits outstanding performance in detecting deepfake audio, achieving an audio accuracy of 0.9508464693536824 and an overall confidence level of 0.9649833786365379.

6. RESULT

The trained model shows the most accurate result in Fig.(4) , Fig.(5), Fig.(6), the deepfake audio detection with audio accuracy: 0.9508464693536824, overall confidence: 0.9649833786365379

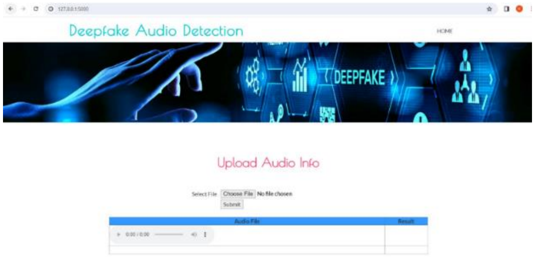
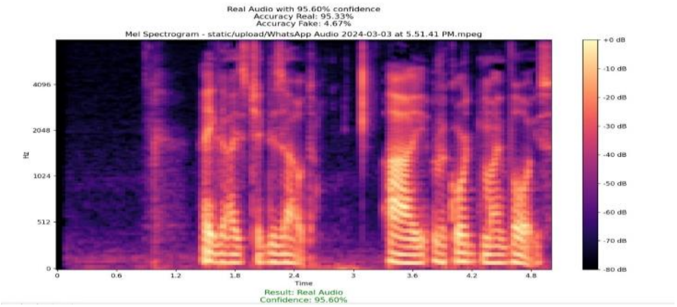
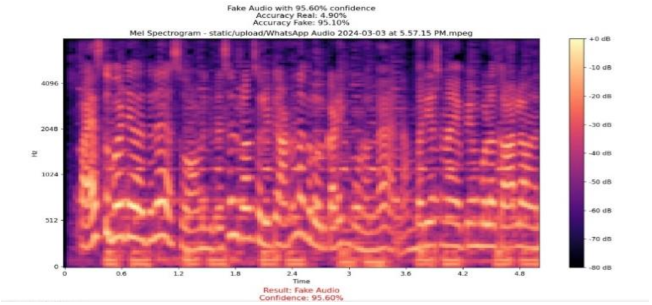


Fig.4. Audio Upload Page


Real Audio with 95.60% confidence
Accuracy Real: 95.33%
Accuracy Fake: 4.67%
Mel Spectrogram - static/upload/WhatsApp Audio 2024-03-03 at 5.51.41 PM.mpeg

Result: Real Audio
Confidence: 95.60%

Fig.5. Bonafide audio result


Fake Audio with 95.60% confidence
Accuracy Real: 4.90%
Accuracy Fake: 95.10%
Mel Spectrogram - static/upload/WhatsApp Audio 2024-03-03 at 5.57.15 PM.mpeg

Result: Fake Audio
Confidence: 95.60%

Fig.6. Spoof audio result

7. CONCLUSION

The detection of deepfake audio presents a formidable challenge within the realm of digital forensics and media integrity verification. As artificial intelligence and machine learning continue to evolve, deepfake technologies are becoming increasingly sophisticated and widespread. Despite these complexities, ongoing research and development endeavors are focused on devising effective strategies to detect and mitigate the dissemination of altered audio content. A range of techniques is under exploration, including spectral pattern analysis, identification of speech characteristic inconsistencies, and leveraging blockchain technology for immutable verification. These approaches hold promise in addressing the issue by enhancing the ability to identify manipulated audio materials. It's important to note that while these detection methods offer a level of defense against deepfake audio proliferation, they are not foolproof and require continual improvement to keep pace with evolving manipulation tactics.

REFERENCES

[1] S. Waseem and Mhassen Elnour Elneel Dalam, "Deep Fake on Face and Expression Swap: A Review", *IEEE Access*, Vol. 13, pp. 1-15, 2023.

[2] Y. Patel, Sudeep Tanwar, Pronaya Bhattacharya, Rajesh Gupta and Turki Alsuwian, "Deepfake Audio Detection via MFCC Features using Machine Learning", *IEEE Access*, Vol. 11, pp. 22081-22095, 2022.

[3] Ahmed Abbasi, 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*, Vol. 10, pp. 38885-38894, 2022.

[4] V. Phani and Krishna Deep, "Fake Detection using LSTM and RESNEXT", *Journal of Engineering Sciences*, Vol. 13, No. 7, pp. 1-11, 2022.

[5] Pramod Dhamdhere, "Semantic Trademark Retrieval System based on Conceptual Similarity of Text with Leveraging Histogram Computation for Images to Reduce Trademark Infringement", *Webology*, Vol. 18, No. 5, pp. 1-9, 2021.

[6] J. Khochare, C. Joshi, B. Yenarkar, S. Suratkar and F. Kazi, "A Deep Learning Framework for Audio Deepfake Detection", *Arabian Journal for Science and Engineering*, Vol. 77, pp. 1-12, 2021.

[7] D. Cozzolino, M. Niebner and L. Verdoliva, "Audio-Visual Person-of-Interest Deepfake Detection", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 1-5, 2022.

[8] Yi Yangyan, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma and Chenglong Wang, "The First Audio Deep Synthesis Detection Challenge", *Proceedings of International Conference on Acoustics Speech and Signal Processing,* pp. 1-9, 2022.

[9] Y. Gao, T. Vuong, M. Elyasi, G. Bharaj and R. Singh, "Generalized Spoofing Detection Inspired from Audio Generation Artifacts", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 1-5, 2021.

[10] R. Yamamoto, E. Song, and J. Kim, "Parallel Wavegan: A Fast Waveform Generation Model based on Generative Adversarial Networks with Multi-Resolution Spectrogram", *Proceedings of International Conference on Machine and Deep Learning*, pp. 1112-1119, 2020.