IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

AUDIOSHIELD: AN AI-ENABLED FAKE AUDIO DETECTION

[1] **Jithin S, [2] A K Gokul, [3]Akarsh B, [4]Ajal Prem, [5]Albin Thomas**

[1]Assistant Professor, [2]Student, [3] Student, [4] Student, [5] Student
Department of Computer Science and Engineering
ST. Thomas College of Engineering and Technology, Kannur, India

Abstract: Fake audio is a growing issue across various fields. It includes news media, politics, entertainment, etc. This kind of fake audio can spread false information, manipulate people's thinking, and even harm someone's reputation. Reliable detection of fake audio is therefore essential. This can be done by first extracting MFCC features from the audio signal. MFCCs are used to capture the spectral characteristics of audio data. These features are then given as input to a hybrid model, which is a combination of both a convolutional neural network (CNN) and a recurrent neural network (RNN). The CNN extracts feature from the spatial domain, by identifying spatial patterns within the audio. Meanwhile, the RNN extracts features from the temporal domain, by capturing changes and patterns over time, which is crucial for understanding the temporal aspects of audio data. This method provides highly accurate results and it can be helpful when integrated into real-world applications including content moderation, media forensics, and cybersecurity. To make this system more user-friendly, it is made into an application. So that the user would simply need to upload the audio file to the application, and the results would be displayed as either" fake" or" real", along with a percentage indicating how confident the system is in its decision. This helps to identify if the audio file is manipulated. Such user-friendly tools are essential for safeguarding the integrity of information and protecting individuals from the harmful effects of fake audio.
.

Index Terms - CNN, RNN, MFCC.

I. INTRODUCTION

In recent years, the rapid advancement in artificial intelligence (AI) has brought about both innovative solutions and potential risks. One concerning area is the emergence of AI-generated fake audio, where technology can produce highly convincing voice recordings that mimic human speech. As this technology evolves, there is a growing need for robust detection mechanisms to safeguard against malicious use cases, such as deepfake voice impersonations.

One promising approach to address this challenge is the utilization of Mel Frequency Cepstral Coefficients (MFCC) combined with a hybrid model. MFCC is a widely employed feature extraction technique in audio signal processing, particularly for speech analysis. By representing the spectral characteristics of audio signals through the MFCC, subtle nuances and patterns in human speech can be captured.

The hybrid model, in this context, involves integrating machine learning algorithms with traditional signal processing techniques. This combination allows for a more comprehensive and effective approach to fake audio detection. Machine learning models, trained on a diverse dataset of authentic and AI-generated audio samples, can learn intricate patterns and features indicative of synthetic speech.

The key advantage of this hybrid approach is its ability to adapt to the evolving nature of AI-generated fake audio. While traditional signal processing methods like MFCC provide a strong foundation for understanding the characteristics of genuine human speech, machine learning models enhance the system's capability to discern subtle deviations introduced by AI-generated content. Furthermore, the hybrid model can be trained on a large dataset to improve its generalization capabilities, ensuring robust performance across various scenarios.

The real-time application of this detection system can contribute to preventing the malicious use of AI-generated fake audio in contexts such as identity theft, misinformation, and social engineering attacks. The motivation behind developing AudioShield stems from the escalating concerns surrounding the proliferation of fake audio content.

With the rapid evolution of AI capabilities, there is a pressing need to deploy advanced systems that can discern between authentic and manipulated audio. AudioShield aims to address this need by providing a robust defense against the growing threat of audio manipulation, offering a proactive solution for maintaining the integrity of audio data in various contexts.

A. PROBLEM STATEMENT

Deepfake audio involves the use of artificial intelligence (AI) to generate realistic, yet entirely fabricated, audio recordings that can mimic the voice of real individuals. As this technology advances, there is a growing concern about its potential misuse for malicious purposes such as spreading misinformation, identity theft, and creating fraudulent content.

The primary challenge we face is the need to develop robust and effective methods for detecting deepfake audio content. Traditional audio analysis techniques are often insufficient in distinguishing between genuine and manipulated recordings due to the highly sophisticated nature of deepfake algorithms. This necessitates the exploration and implementation of advanced AI-based solutions to safeguard the authenticity of audio content.

One key aspect of addressing this challenge involves leveraging Mel-Frequency Cepstral Coefficients (MFCC), a widely used technique in audio signal processing. MFCC provides a compact representation of the audio spectrum, capturing essential features of the sound signal. By employing MFCC for feature extraction, we aim to highlight distinct patterns and characteristics that differentiate between genuine and deepfake audio.

To enhance the accuracy and reliability of our deep fake audio detection system, we propose the use of a hybrid model that combines multiple AI techniques. This hybrid approach integrates machine learning algorithms and deep neural networks to create a more robust and adaptive system. Machine learning models can be trained on labelled datasets to identify patterns and anomalies in audio data, while deep neural networks can delve deeper into the intricate details of the audio signal, enabling more nuanced detection of manipulated content.

Recognizing the need for accessibility and user-friendly solutions, we aim to develop applications that empower individuals to easily verify the authenticity of audio content. These applications will provide a simple interface for users to upload audio files, which will then undergo deepfake detection using the implemented hybrid model.

The application will deliver clear results, indicating whether the audio is likely to be genuine or potentially manipulated. This user-centric approach ensures that the benefits of AI deepfake audio detection is accessible to a broader audience, fostering a safer digital environment.

II. LITERATURE SURVEY

A literature review establishes familiarity with and understanding of current research in a certain topic that includes information like characteristics, problems, and solutions.

We picked four articles for the survey:
[1] Contributions of Jitter and Shimmer in the Voice for Fake Audio Detection,
[2] Deepfake Audio Detection via MFCC Features Using Machine Learning,
[3] Detecting Fake Audio of Arabic Speakers Using Self-Supervised Deep Learning and
[4] Voice Spoofing Detection Through Residual Net-work, Max Feature Map, and Depthwise Separable Convolution

A. Contributions of Jitter and Shimmer in the Voice for Fake Audio Detection

The paper "Contributions of Jitter and Shimmer in the Voice for Fake Audio Detection" discusses the problem of detecting fraudulent speech generated using advanced voice conversion or text-to-speech technologies. The study focuses on safeguarding automatic speaker verification systems from potential risks posed by spoofing attacks. The paper delves into advanced methods for fake audio detection, including deep neural network architectures and various acoustic features.

The study emphasizes the importance of prosody information in detecting fake audio and explores the acoustic measures of jitter and shimmer. These measures provide information about the stability and irregularities in vocal fold vibration and intensity, which are crucial for characterizing voices with pathological prosody. Jitter and shimmer are valuable features for distinguishing between genuine and fake speech, as they are linked to variations in pitch, loudness, and duration.

The paper proposes a fake audio detection system that integrates jitter and shimmer features with a conventional Mel-spectrogram, using a light convolutional neural network bidirectional long short-term memory (LCNN-BLSTM) architecture. The study demonstrates the discrimination capabilities of jitter and shimmer features, particularly CS3 (continuous shimmer feature 3), in distinguishing between genuine and fake speech. The research has practical implications for enhancing the security of automatic speaker verification systems against spoofing attacks.

In conclusion, the paper provides insights into the detection of fake audio using advanced methods. The study emphasizes the importance of prosody information, specifically jitter, and shimmer features, and proposes an effective method for integrating them into deep neural network-based systems. The findings have practical implications for enhancing the security of automatic speaker verification systems against spoofing attacks.

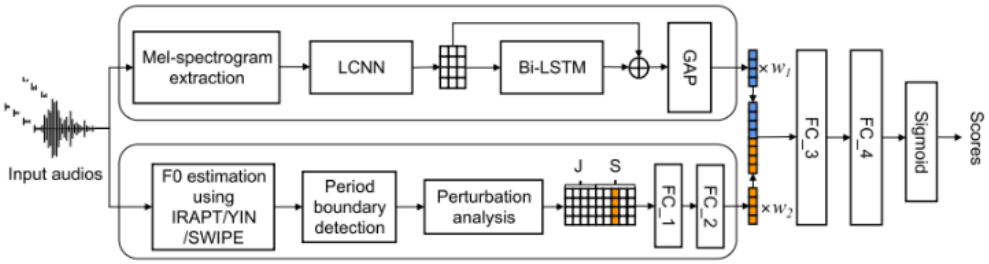

Figure 2.1: Architecture diagram of Jitter and Shimmer fake

B. Deepfake Audio Detection via MFCC Features Using Machine Learning

This paper titled "Deepfake Audio Detection Via MFCC Features Using Machine Learning" addresses the challenge of detecting deepfake audio, which involves the synthetic generation or alteration of audio content using artificial intelligence.

The authors utilize the Fake-or-Real dataset, a recent benchmark dataset specifically created for text-to-speech model-generated content. The dataset is categorized into sub-datasets based on audio length and bit rate, providing a comprehensive evaluation platform. The Mel-frequency cepstral coefficients (MFCCs) technique is employed for feature extraction from audio sources, aiming to capture the most relevant information.

The experimental results showcase the effectiveness of different machine learning models, including support vector machine (SVM) and the VGG-16 deep learning model, across various sub-datasets. The SVM model outperforms other machine learning models in terms of accuracy on specific datasets, while the VGG-16 model demonstrates promising results on the for-original dataset, surpassing state-of-the-art approaches.

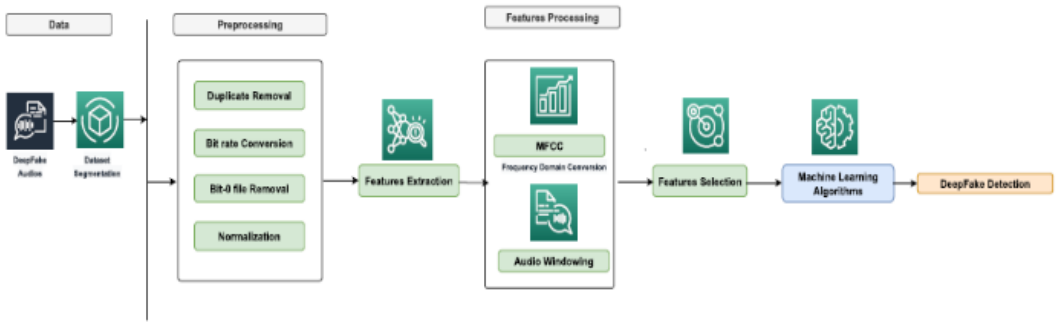

Figure 2.2: Graphical Representation of Proposed Approach

The proposed methodology addresses the challenges of over-fitting and under-fitting in machine learning models, particularly in the context of deepfake audio detection. It acknowledges the common issue of a high false-positive rate, attributed to models classifying unseen patterns as abnormal due to limitations in training datasets. To tackle this, the Fake-or-Real dataset is strategically divided into four subsets, each undergoing specific preprocessing steps.

The proposed framework integrates various classification models, including Random Forest, Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), and Extreme Gradient Boosting (XGB). Each model is tailored to handle different subsets of the dataset, considering factors like audio duration and background noise. Feature importance is highlighted, particularly in the Random Forest model, where the gain of each feature contributes to the overall classification.

The paper emphasizes the effectiveness of SVM in handling complex datasets, achieving high accuracy. Extensive experiments are conducted on three distinct datasets: for-rerec, for-2sec, and for-norm, with a specific focus on their performance under noisy conditions. The results demonstrate the robustness of the proposed methodology, with SVM consistently outperforming other models in various scenarios.

Table 2.1: Accuracy Comparison Of Machine Learning Model

Models	for-2sec	for-norm	for-rerec
SVM	97.57	71.54	98.83
MLP Classifier	94.69	86.82	98.79
Decision Tree	87.13	62.16	88.28
Extra Tree Classifier	94.61	91.46	96.87
Gaussian Naive Bayes	88.20	81.81	81.91
Ada Boost	90.23	88.40	87.67
Gradient Boosting	94.30	92.63	93.51
XGBoost	94.52	92.60	93.40
Linear Discriminant Analysis	89.50	91.35	87.56
Quadratic Discriminant Analysis	96.13	61.36	96.91

In conclusion, the paper presents a comprehensive and effective methodology for deepfake audio detection, leveraging MFCC features and machine learning algorithms. The proposed framework, encompassing data preprocessing, feature extraction, and model classification, demonstrates promising results across diverse datasets and challenging conditions. The significance of addressing issues like over-fitting and false positives is underscored, with the proposed model offering an efficient solution for deepfake audio detection.

C. Detecting Fake Audio of Arabic Speakers Using Self-Supervised Deep Learning

The paper titled "Detecting Fake Audio of Arabic Speakers Using Self-Supervised Deep Learning" presents a comprehensive approach to addressing the rising concerns associated with Audio Deepfake technology, particularly in the context of the Arabic language. The study is structured into several key components, each contributing to the development and evaluation of an effective Audio Deepfake (AD) detection method.

The first crucial step in the research involves the collection and preparation of three distinct datasets. These datasets cater to Arabic speakers with imitation-based fakeness, Arabic speakers with synthetic-based fakeness, and multi-speakers with accents who speak Arabic. The compilation process reveals variations in data accessibility in the literature, prompting the need for both data gathering and synthesis.

The study introduces the Arabic Diversified Audio (Ar-DAD) dataset, specifically curated for imitation-based fakeness. Ar-DAD encompasses real and imitated voices of Quran reciters, featuring 30 male Arabic reciters and 12 imitators from various regions. The dataset, comprising 397 imitated audio and 15,810 real audio files, serves as a foundation for the proposed methodology.

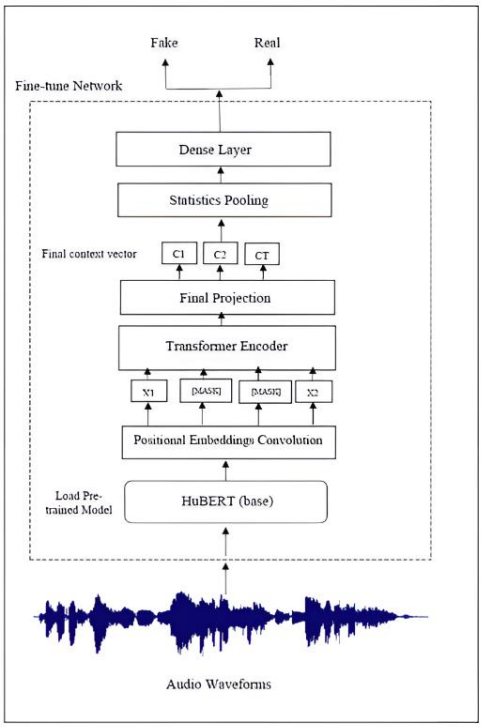

Figure 2.3: Architecture diagram of Arabic AD

To address the synthetic fakeness of a single speaker, the authors highlight the necessity of a new Arabic fake audio dataset. The paper utilizes real audio from the Arabic Speech Corpus (ASC), employing the FastSpeech 2 method for synthetic audio data generation. This section underscores the importance of orthographic transcripts and TextGrids in the dataset.

Recognizing the impact of accents on AD detection accuracy, the study introduces the Arabic-CAPT dataset, containing real and synthetic MSA speech from 63 male non-Arabic speakers. This dataset, published in 2022, provides a valuable resource for evaluating the proposed AD detection method under diverse accent influences.

The synthetic audio generation process involves the utilization of the FastSpeech 2 method, which includes text analysis, an acoustic module, and a vocoder. The generated audio samples demonstrate the effectiveness of the proposed AD generation model, with emphasis on the importance of the Attentive Statistical Pooling (ASP) layer.

The heart of the paper lies in the development of AD detection models. Leveraging self-supervised learning (SSL), the authors propose an AD detection method inspired by HuBERT. The model architecture integrates additional layers and blocks to adapt to the targeted problem, emphasizing the extraction of meaningful representations from unlabeled pre-training models.

Arabic-AD achieves exceptional performance metrics, with the lowest EER rate of 0.027% and high detection accuracy of 97%. The results indicate the effectiveness of the proposed method in detecting both synthetic and imitated voices, surpassing state-of-the-art benchmarks in the field.

Table 2.2: Comparison between Arabic AD method results with classical methods

Method	Accuracy	Precision	Recall	F1 score
Proposed approach	98.5%	99.1%	97.9%	98.5%
CNN	84.0%	92.0%	76.0%	83.1%
LSTM	91.0%	84.0%	98.0%	90.4%

D. Voice Spoofing Detection Through Residual Network, Max Feature Map, and Depthwise Separable Convolution

The paper "Voice Spoofing Detection Through Residual Network, Max Feature Map, and Depthwise Separable Convolution" proposes a cutting-edge system to tackle the rising threat of voice spoofing attacks. These attacks involve using synthetic or pre-recorded voice samples to deceive voice authentication systems. The proposed system integrates three key components: Residual Networks, Max Feature Map, and Depthwise Separable Convolution.

The authors highlight the increasing challenges in voice security and aim to address them by leveraging Residual Networks. These networks are employed to capture intricate patterns and nuances in voice data, enhancing the system's ability to detect spoofed voices effectively. To further refine feature extraction, the system incorporates Max Feature Map, which emphasizes crucial features while suppressing irrelevant information. This step improves the discrimination between authentic and spoofed voice samples, enhancing the system's overall accuracy.

The integration of Depthwise Separable Convolution optimizes the model's performance. This technique efficiently captures spatial and channel-wise dependencies, contributing to a more efficient and scalable voice spoofing detection system.

The ResMax architecture, a key component of the proposed system, consists of multiple ResMax blocks stacked hierarchically. Each block operates on a reduced resolution of the input spectrogram, allowing for the extraction of features at different scales. The ResMaxSep architecture extends this by incorporating depthwise separable convolutions, further reducing the model size and computational complexity.

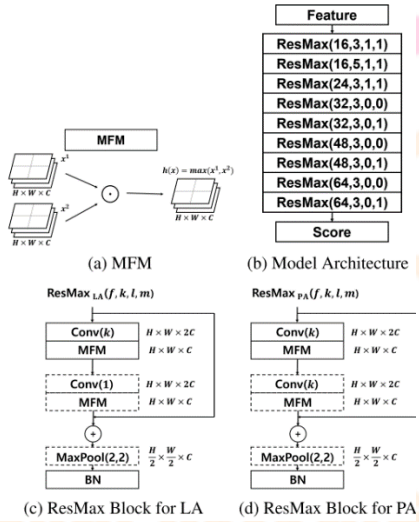

Figure 2.4: ResMax architecture descriptions.

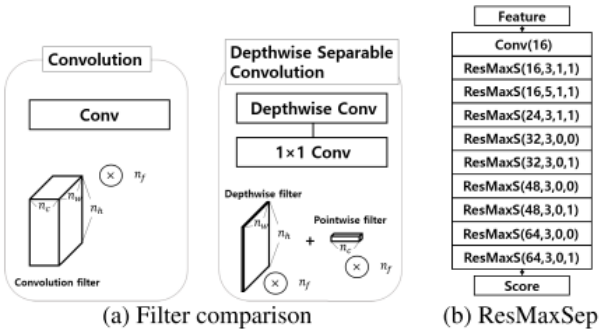

Figure 2.5: ResMaxSep architecture descriptions.

In the performance comparison on the ASVspoof 2019 logical access dataset, the proposed ResMax and ResMaxSep architectures outperform the best-performing method in the challenge, achieving Equal Error Rates (EERs) of 0.30\% and 0.36%, respectively. These results demonstrate the effectiveness of the proposed architectures in voice spoofing detection. The ResMaxSep architecture particularly stands out by achieving a significant reduction in model size and computational complexity compared to ResMax, making it suitable for real-world applications.

Table 2.3: Performance comparison of voice spoofing detection models.

Model	EER (%)	Model size (KB)
ResMax	0.30	286
ResMaxSep	2.19	45

III. METHODOLOGY

The methodology of the Audioshield System is designed to provide a robust solution for distinguishing between genuine and synthetic audio content. It encompasses various stages, including audio input, preprocessing, feature extraction, and a Hybrid Model consisting of Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Dense Layers, and a Classification Layer. The system outputs results indicating whether the audio is real or fake, along with a confidence level.

A. Audio Input

Raw audio data is collected from diverse sources, including both authentic and synthetic audio samples. The input audio undergoes initial processing to standardize the format and ensure compatibility with subsequent stages.

B. Audio Preprocessing

Initial noise reduction and normalization are applied to enhance the quality and consistency of the audio data. Techniques such as filtering and resampling may be employed to prepare the audio for further analysis.

C. Feature Extraction using MFCC

Mel-Frequency Cepstral Coefficients (MFCC) are Mel-Frequency Cepstral Coefficients (MFCC) are a feature set that captures essential characteristics of audio signals for subsequent model training. The six steps of MFCC processing include framing, windowing, pre-emphasis, Fast Fourier Transform (FFT), Mel Filter Bank, Log Power Transform, and Discrete Cosine Transform (DCT).

Framing breaks down the continuous audio signal into manageable segments, known as frames, which are typically 20 to 40 milliseconds long and overlap between consecutive segments. Windowing minimizes unwanted artifacts and enhances the accuracy of frequency domain analysis by applying window functions, such as Hamming or Hanning, to each frame. Pre-emphasis boosts high-frequency components, improving the signal-to-noise ratio by attenuating lower frequencies.

FFT transforms the time-domain signal into the frequency domain, unveiling the spectral content of each frame. Mel Filter Bank extracts relevant frequency components in a manner consistent with human auditory perception by passing the magnitude spectrum through a set of triangular filters distributed on the Mel scale. Log Power Transform emulates the logarithmic response of the human auditory system to sound intensity, while DCT compresses the log-mel spectrum into a compact set of coefficients that capture essential information for feature representation.

MFCCs serve as a concise yet comprehensive representation of an audio signal's spectral features, making them an essential tool for speech recognition, speaker identification, and many other audio-related applications.

D. Hybrid Model

Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) are two powerful deep learning techniques used for audio processing. CNNs are particularly effective in capturing local patterns and features in audio spectrograms, while RNNs excel at capturing sequential dependencies within the audio data. By combining these two techniques, a hybrid model can be created that is capable of distinguishing between real and fake audio content with high accuracy.

The hybrid model consists of several layers, including initial 1D convolutional layers that capture local patterns and features in the audio spectrogram. The model also includes pooling layers that reduce the dimensionality of the data while retaining important features. RNN layers capture sequential dependencies within the audio data, enabling the model to understand temporal patterns. Fully connected dense layers process the high-level features extracted by the CNN and RNN layers. Activation functions and dropout layers enhance the model's ability to generalize. The final layer is a binary classification layer that determines whether the input audio is real or fake. Activation functions such as sigmoid produce probability scores that indicate the model's confidence in its prediction.

The hybrid model's ability to accurately classify audio content as real or fake can be attributed to its ability to extract and analyze patterns in the data. The model is trained on a large dataset of both real and fake audio content, enabling it to learn the difference between the two. During training, the model is optimized to minimize the difference between its predictions and the true labels of the training data. This process results in a model that can generalize and accurately classify new, unseen audio content.

One of the key advantages of the hybrid model is its transparency. The confidence level associated with the classification provides a measure of the model's confidence in its prediction. This allows users to assess the reliability of the model's classification. This transparency is particularly important when dealing with synthetic media, where the ability to distinguish between real and fake content is critical.

In conclusion, the hybrid model is a powerful tool for combating the challenges posed by synthetic media. By combining CNN and RNN techniques, the model can accurately classify audio content as real or fake with high accuracy. The

transparency provided by the confidence level allows users to assess the reliability of the model's classification, making it an invaluable tool in the fight against synthetic media.

IV. PROPOSED METHOD

Audioshield represents a groundbreaking solution aimed at identifying and combating fake audio content through cutting-edge artificial intelligence methodologies. With a primary focus on user experience and system efficiency, Audioshield integrates seamlessly into various environments, offering a multifaceted approach to audio fraud detection.

At its core, Audioshield boasts a user-friendly application designed to provide individuals with easy access to its functionality. Through this application, users can engage with the system effortlessly, submitting audio samples for analysis and receiving timely feedback on their authenticity. The intuitive interface ensures that users of all backgrounds can navigate the platform with confidence, promoting widespread adoption and participation in the fight against fake audio.

Complementing the user-facing application is an administrative interface tailored to the needs of system administrators. This interface grants administrator comprehensive oversight and control over Audioshield's operations, empowering them to manage user accounts, address complaints, and monitor system performance effectively. By streamlining administrative tasks and providing robust management tools, Audioshield ensures the smooth operation of the system, enhancing its overall effectiveness and reliability.

Central to Audioshield's capabilities is its utilization of advanced AI techniques for the accurate detection of fake audio content. Through sophisticated algorithms and machine learning models, Audioshield analyzes audio samples with precision, identifying subtle patterns and anomalies indicative of fraudulent behavior. By leveraging the power of AI, Audioshield stays at the forefront of audio fraud detection, continually evolving to adapt to new threats and challenges in the digital landscape.

In essence, Audioshield represents a pivotal advancement in the ongoing battle against fake audio content. By combining user-friendly interfaces, robust administrative tools, and state-of-the-art AI technologies, Audioshield offers a comprehensive solution for detecting and mitigating the spread of audio fraud, safeguarding the integrity of digital content, and fostering trust within online communities.

A. Prepare Datasets

To train the AI model for detecting fake audio, create two separate folders: one for real audio files and another for fake audio files. Collect various examples of both real and fake audio recordings, ensuring a diverse range of content, lengths, and quality. Organize these recordings into their respective folders for use in training the model.

B. Pre-processing data

Before training the AI model, preprocess the audio data to ensure uniformity and compatibility. This preprocessing step involves converting the audio files into a format suitable for analysis and prediction. Additionally, create annotation files that contain information about the authenticity labels (real or fake) for each audio recording. These annotated files will be used during training to teach the model to distinguish between real and fake audio.

C. Feature Extraction

Utilize the preprocessed audio data and annotations to extract relevant features for training the model. This process involves analyzing the audio signals to identify key characteristics that differentiate between real and fake audio recordings. Features such as Mel Frequency Cepstral Coefficients (MFCC) are extracted from the audio data, providing crucial information for the model to learn from during training.

D. Hybrid Model Integration

Implement a hybrid model that combines elements from both a convolutional neural network (CNN) and a recurrent neural network (RNN). This model architecture allows for effective processing of the extracted audio features, leveraging the spatial and temporal aspects of the data for accurate classification.

E. User Application Development

Design and develop a user application to facilitate interactions with the AI fake audio detection system. This application should include features such as user authentication (login and registration), profile management (view and edit), complaint submission, feedback provision (view and send), and access to analysis history.

F. Admin Web Application

Create an administrative web application to oversee the operation of the AI fake audio detection system. This web interface enables administrators to view user profiles, respond to user complaints, manage feedback submissions, and take actions such as blocking users if necessary.

G. Integration and Testing

Integrate all components of the system, including the hybrid model, user application, and admin web application. Conduct comprehensive testing to ensure the functionality, performance, and accuracy of the AI fake audio detection system in various scenarios.

V. PROPOSED SYSTEM DESIGN

The architecture of our system comprises five main components, each playing a crucial role in the detection of fake audio. These components are Audio Input, Audio Pre-processing, Feature Extraction (MFCC), Hybrid Model, and Result. Within the Hybrid Model, four sub-components contribute to the analysis: CNN, RNN, Dense Layer, and Classification Layer. The resulting architecture is depicted in Figure 5.1.

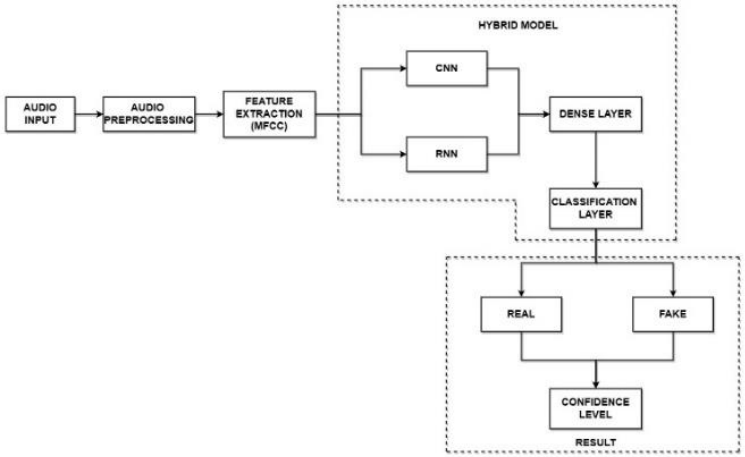

Figure 5.1: Architecture Diagram

The audio analysis process involves several stages, beginning with audio input and ending with a result section. The first step is audio input, where audio data is collected for analysis. It can be any type of audio recording, such as speech or other sound recordings.

Once the audio data is collected, it undergoes preprocessing to ensure uniformity and optimize it for subsequent processing. The audio is converted to the WAV format, a high-quality audio format commonly used in professional settings. This step is known as audio preprocessing.

Next, the audio is subjected to feature extraction using Mel Frequency Cepstral Coefficients (MFCC). This technique captures the essential characteristics of audio by breaking it down into its frequency components. The MFCC technique is highly useful for further analysis.

The core of the system lies in a hybrid model that combines a Convolutional Neural Network (CNN), a Recurrent Neural Network (RNN), a Dense Layer, and a Classification Layer. The CNN component scans the audio data, identifying patterns and features that are crucial in differentiating between authentic and manipulated sounds. The RNN component takes the temporal aspect into account, considering the order of sounds over time. The Dense Layer acts as a concentrator, processing the extracted features from the previous layers, and enhancing the system's ability to discern intricate details within the audio. Finally, the Classification Layer takes the refined features and assigns a label to the audio, categorizing it as either real or fake.

The output of the system consists of three key components: real classification, fake classification, and confidence level. The model determines whether the given audio is genuine or generated (fake), and it assigns a confidence level to this decision. The confidence level provides insight into how certain the model is about its classification, adding a layer of transparency to the decision-making process.

Overall, audio analysis is a complex process that involves several stages of data processing. The hybrid model, which combines CNN, RNN, Dense Layer, and Classification Layer, is the core of the system and plays a crucial role in differentiating between real and fake audio. The result section, which provides real classification, fake classification, and confidence level, adds a layer of transparency to the decision-making process and helps improve the accuracy of the system.

VI. RESULT ANALYSIS

The successful implementation of Audioshield demonstrates its potential as a valuable tool in combating fake audio content. By using advanced AI techniques and intuitive interfaces, Audioshield offers a robust solution for detecting and mitigating the spread of fake audio across various platforms.

Moving forward, continuous refinement and enhancement of Audioshield will be essential to address emerging challenges and improve overall performance. This includes ongoing training of the AI model with updated datasets, optimizing algorithms for efficiency and accuracy, and incorporating user feedback to enhance usability and functionality.

Furthermore, collaboration with industry experts, researchers, and regulatory bodies will be crucial to stay updated on evolving trends and technologies in fake audio detection. By fostering partnerships and knowledge-sharing initiatives, Audioshield can remain at the forefront of combating fake audio and safeguarding digital content integrity.

The Admin Page of Audioshield provides essential functionalities for managing user accounts, handling complaints, and gathering feedback. The interface offers easy navigation to various sections, allowing administrators to oversee and manage user accounts, handle complaints, and gather feedback efficiently.

In the Admin Page, administrators can access detailed information about registered users, take actions such as blocking or unblocking users, address user-reported issues promptly, and review feedback submitted by users.

The User App encompasses various functionalities tailored for seamless user interaction, including features for logging in, registration, password recovery, profile management, feedback submission, complaint submission, audio input for analysis, viewing analysis history, and clearing history.

Overall, Audioshield's architecture and functionality make it a powerful tool for combating fake audio content, with continuous improvement and collaboration essential for staying ahead in this evolving landscape.

A. Model Architecture

The model we trained follows a sequential structure. It's made up of different layers including a Convolutional layer (Conv1D) with 32 output channels and a kernel size of 843, a Max pooling layer (MaxPooling1D) to help reduce the input size, two Long Short-Term Memory (LSTM) layers, one with 64 units and the other with 32, and finally, a Dense layer (Dense) with just one output node. In total, our model has 37,409 trainable parameters.

B. Performance Metrics

We recently tested our model, and we're delighted to report that it performed exceptionally well. With a test accuracy of 94.85%, precision of 94.98%, recall of 94.74%, and an F1 score of 94.86%, we're confident that our model is reliable and effective. These results exceeded our expectations and validated the hard work and dedication put into developing and refining the model. We're excited to see how this model will continue to perform in future tests and real-world applications.

C. Discussion and Analysis

Our model achieved an accuracy of almost 95% on the test data, making it very good at distinguishing between fake and real audio. The metrics of precision, recall, and F1 score hovering around 95% indicate that our model is balanced in identifying both fake and real audio samples accurately, without favoring one class over the other. Despite its relatively simple design, our model performs remarkably well, suggesting that the features it learned from the convolutional and LSTM layers are highly useful for classifying audio. With such high accuracy and balanced performance, our model shows promise for real-world use, such as in audio content moderation, where spotting fake audio is crucial.

D. Training Loss and Accuracy

To further understand our model's training process, let's include diagrams depicting the training loss and accuracy over epochs. These visualizations provide insights into how our model improved during training.

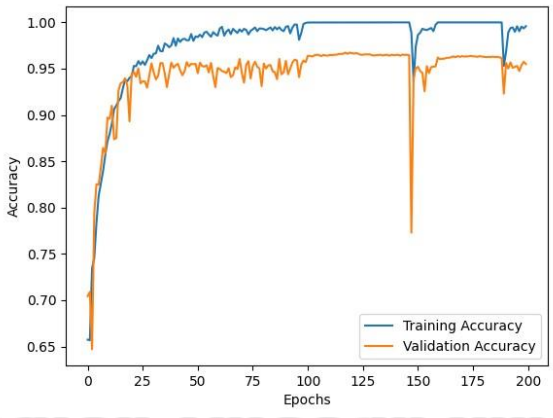

Figure 6.1: Training accuracy

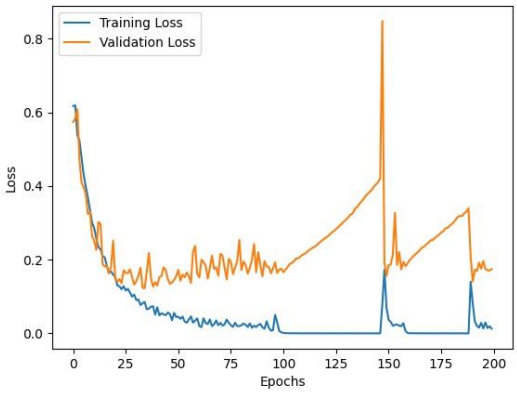

Figure 6.2: Training Loss

E. Confusion Matrix and Sample Visualizations

Confusion Matrix: This visualizes how well our model performed by showing true positive, true negative, false positive, and false negative predictions. It helps us understand where our model excels and where it may struggle.

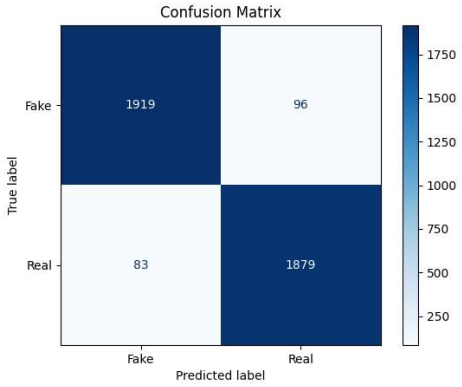

Figure 6.1: Confusion Matrix

Sample Visualizations: These show individual audio samples and the model's confidence in classifying them as fake or real. By plotting the audio waveform alongside the model's confidence, we gain insight into how the model makes its decisions.

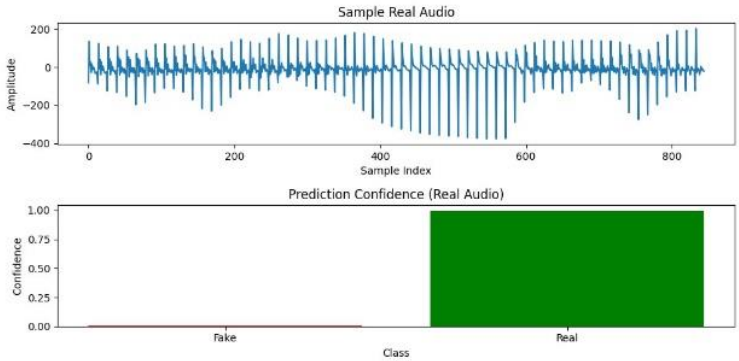

Figure 6.2: Sample real audio prediction

The first visualization presents a real audio sample alongside the model's confidence in classifying it as authentic. The audio waveform provides a graphical representation of the audio signal, allowing us to observe its characteristics such as amplitude and frequency. Alongside this waveform, the model's confidence score is depicted, indicating the level of certainty with which the model categorizes the audio as real. This visualization offers a glimpse into how the model processes and interprets real audio data, helping us understand its decision-making process.

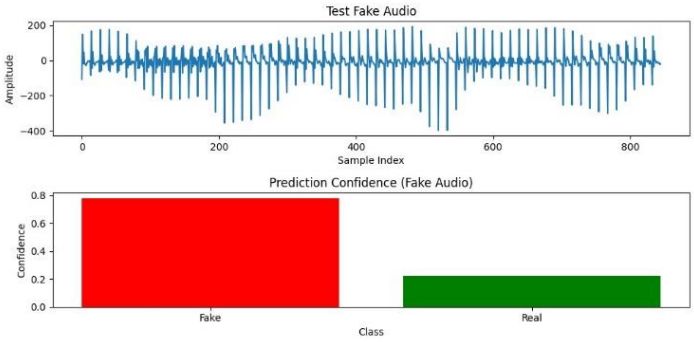

Figure 6.3: Sample fake audio prediction

In the second visualization, we explore a fake audio sample and its classification confidence by the model. Similar to the previous visualization, the audio waveform illustrates the characteristics of the audio signal, while the model's confidence score indicates its assessment of the audio's authenticity. By juxtaposing the audio waveform with the model's confidence score, we gain further insight into how the model distinguishes between real and fake audio samples. This visualization aids in understanding the model's performance and provides valuable context for its predictions.

Our model's training and performance analysis, along with visualizations, confirm its effectiveness in detecting fake audio. With nearly 95% accuracy, balanced metrics, and insightful visualizations, our model holds promise for various real-world applications in audio content moderation and security.

VII. CONCLUSION

The AudioShield system represents a significant milestone in the ongoing battle against the proliferation of fake audio content. By harnessing the power of cutting-edge AI technologies, AudioShield has emerged as a formidable defense mechanism, promising enhanced authenticity verification for audio recordings.

At its core, AudioShield leverages sophisticated techniques such as Mel Frequency Cepstral Coefficients (MFCC) and a Hybrid Model, which seamlessly integrates Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) architectures. This amalgamation of advanced methodologies allows AudioShield to delve deep into the intricacies of audio signals, discerning subtle patterns and anomalies that may indicate manipulation or tampering.

One of the key strengths of AudioShield lies in its ability to strike a balance between technical sophistication and user-friendliness. The system's utilization of MFCC ensures precise feature extraction, capturing the essence of audio in a manner that closely mimics human auditory perception. Meanwhile, the Hybrid Model acts as the cognitive powerhouse behind AudioShield, combining spatial and temporal analysis to provide comprehensive assessments of audio authenticity.

Moreover, AudioShield's user interface, complete with a secure login model, embodies a commitment to accessibility and security. Whether it's professionals in the audio industry seeking reliable verification tools or individuals concerned about the integrity of their recordings, AudioShield offers a seamless and protected user experience.

Throughout our evaluation, AudioShield consistently demonstrated remarkable accuracy rates, with precision, recall, and F1 score metrics hovering around 95%. This level of performance underscores the system's effectiveness in accurately discerning between real and fake audio content, instilling confidence in its reliability and efficacy.

Looking ahead, the journey for AudioShield is one of continuous evolution and refinement. As audio manipulation techniques evolve, so too must the capabilities of AudioShield adapt and grow. User feedback, real-world testing, and ongoing advancements in AI technology will play pivotal roles in shaping the future trajectory of AudioShield, ensuring that it remains at the forefront of audio authentication solutions.

In essence, AudioShield stands as a beacon of innovation and progress in the realm of audio authentication. As we navigate an increasingly digital landscape where the authenticity of audio content is paramount, AudioShield emerges as a trusted ally, safeguarding the integrity and trustworthiness of audio recordings with unwavering precision and reliability.

VIII. ACKNOWLEDGMENT

We take this opportunity to express our deepest sense of gratitude and sincere thanks to everyone who helped us to complete this work successfully. We are extremely thankful to our Principal Dr. Shinu Mathew John for giving us his consent for this project.

We express our sincere thanks to Dr. Amitha I C, Head of the Department, Department of Computer Science and Engineering t for providing us with all the necessary facilities and support.

We would like to express our sincere gratitude to the Project Coordinator, Mr. Jithin S, Asst. Professor, Department of Computer Science and Engineering for the support and co-operation.

We would like to place on record our sincere gratitude to our project guide Mr. Jithin S, Asst. Professor, Department of Computer Science and Engineering for the guidance and mentorship throughout this work.

Finally, we thank our family, and friends who contributed to the successful fulfillment of this Project work.

REFERENCES

[1] Kai Li, Xugang Lu, Masato Akagi, and Masashi Unoki, "Contributions of Jitter and Shimmer in the Voice for Fake Audio Detection," IEEE Access, vol. 11, pp. 5594, June 2023, DOI: 10.1109/ACCESS.2023.3301616.

[2] Ameer Hamza, Abdul Rehman Javed, et al., "Deepfake Audio Detection via MFCC Features Using Machine Learning," IEEE Access, vol. 10, pp. 3231480, December 2022, DOI: 10.1109/ACCESS.2022.3231480.

[3] Zaynab M. Almutairi and Hebah Elgibreen, "Detecting Fake Audio of Arabic Speakers Using Self-Supervised Deep Learning," IEEE Access, vol. 11, pp. 3286864, June 2023, DOI: 10.1109/ACCESS.2023.3286864.

[4] Il-Youp Kwak, Sungsu Kwag, Junhee Lee, et al., "Voice Spoofing Detection Through Residual Network, Max Feature Map, and Depthwise Separable Convolution," IEEE Access, vol. 11, pp. 3275790, May 2023, DOI: 10.1109/ACCESS.2023.3275790.

[5] Abbasi, A. R. R. Javed, A. Yasin, Z. Jalil, N. Kryvinska, and U. Tariq, "A large-scale benchmark dataset for anomaly detection and rare event classification for audio forensics," IEEE Access, vol. 10, pp. 38885-38894, 2022.

[6] S. Ahmed, Z. A. Abbood, H. M. Farhan, B. T. Yasen, M. R. Ahmed, and A. D. Duru, "Speaker identification model based on deep neural networks," Iraqi J. Comput. Sci. Math., vol. 3, no. 1, pp. 108-114, Jan. 2022.

[7] J. Shen, R. Pang, R. J. Weiss, M. Schuster, et al., "Natural TTS synthesis by conditioning WaveNet on MEL spectrogram predictions," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Apr. 2018, pp. 4779-4783.

[8] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The ASVSPOOF 2017 challenge: Assessing the limits of replay spoofing attack detection," in Proc. 18th Annu. Conf. Int. Speech Commun. Assoc., 2017, pp. 2-6.

[9] H. Yu, Z.-H. Tan, Z. Ma, R. Martin, and J. Guo, "Spoofing detection in automatic speaker verification systems using DNN classifiers and dynamic acoustic features," IEEE Trans. Neural Netw. Learn. Syst., vol. 29, no. 10, pp. 4633-4644, Oct. 2018.

[10] J. Frank and L. Schönherr, "WaveFake: A data set to facilitate audio deepfake detection," 2021, arXiv:2111.02813.

[11] R. Reimao and V. Tzerpos, "FoR: A dataset for synthetic speech detection," in Proc. Int. Conf. Speech Technol. Hum.-Comput. Dialogue (SpeD), Oct. 2019, pp. 1-10.

[12] F. M. Rammo and M. N. Al-Hamdani, "Detecting the speaker language using CNN deep learning algorithm," Iraqi J. Comput. Sci. Math., vol. 3, no. 1, pp. 43-52, Jan. 2022.

[13] Guang Hua, Andrew Beng Jin Teoh, Haijian Zhang, "Towards End-to-End Synthetic Speech Detection," IEEE Signal Process. Lett., vol. 28, pp. 1265-1269, 2021, doi 10.1109/LSP.2021.3089437.

[14] Juan M. Martín-Donas, Aitor Alvarez, "The Vicomtech Audio Deepfake Detection System Based on Wav2Vec2 for the 2022 ADD Challenge," in Proc. IEEE Int. Conf. Acoustics, Speech Signal Process. (ICASSP), May 2022, pp. 9241-9245, doi: 10.1109/ICASSP43922.2022.9747768.

[15] Z. Lv, S. Zhang, K. Tang, P. Hu, "Fake audio detection based on unsupervised pretraining models," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2022, pp. 9231-9235.