ROBUST HYBRID DEEP LEARNING MODEL FOR VOICE SPOOF DETECTION

A PROJECT REPORT

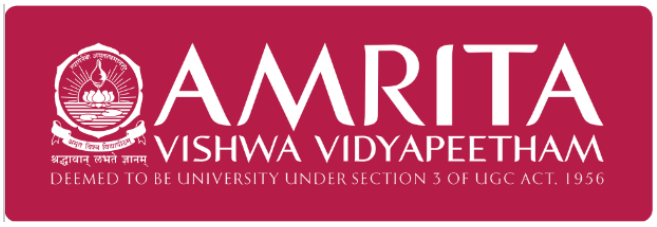
Submitted by

DINGARI JAHNAVI – CH.EN.U4AIE21111
PULATA SANDEEP – CH.EN.U4AIE21137
SASANK SAMI – CH.EN.U4AIE21160

BTECH IN COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE

Under the guidance of
D SASIKALA

Submitted to



AMRITA VISHWA VIDYAPEETHAM
AMRITA SCHOOL OF COMPUTING
CHENNAI – 601103

April 2024

ABSTRACT

This project focuses on the critical task of detecting fake speech, a prevalent form of misinformation with significant societal implications. The primary objective is to develop robust and efficient models capable of discerning between real and fake speech through deep learning techniques. Leveraging Long Short-Term Memory (LSTM) networks and a hybrid model integrating Convolutional Neural Networks (CNN) with LSTM, the project explores various acoustic features extracted from audio data using tools like librosa. Methodologically, the project involves extensive experimentation and evaluation on benchmark datasets such as ASVspoof to assess the efficacy of the proposed models. Data preprocessing steps, including splitting and augmentation, are employed to enhance model performance. Key findings reveal promising results in accurately identifying fake speech, demonstrating the potential of the proposed approach in combating misinformation in digital media platforms. Through waveform visualization, Short-Time Fourier Transform (STFT) analysis, spectrogram examination, and feature extraction techniques like Mel-Frequency Cepstral Coefficients (MFCCs), the project provides insights into the underlying characteristics of authentic and fabricated speech. The research culminates in the development of advanced models that not only advance the field of speech processing technology but also hold practical implications for addressing the spread of misinformation in contemporary digital environments.

Keywords: Fake Speech Detection, LSTM, CNN, MFCC, Acoustic Features.

Iii

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

1.1 BACKGROUND

The emergence of false speech poses a serious threat to public confidence and media credibility in today's digital age. The ease with which phony recordings can now be produced because to the availability of audio editing software is a concern to governmental stability as well as public confidence. The complex and subtle changes applied to audio recordings make it difficult to identify phony speech. Because conventional techniques frequently fail to distinguish between real and artificial audio, researchers are exploring more sophisticated methods like deep learning. By using CNNs and hybrid models that combine LSTM and LSTM, researchers hope to improve the precision and resilience of detection methods. Crucially, creating efficient detection systems requires a grasp of the unique characteristics of speech, such as spectral patterns and grammatical clues.Benchmark datasets, like as ASVspoof, provide offer standardized platforms for rigorous testing and detection algorithm improvement. Experts in cybersecurity, signal processing, and machine learning worked together on this project. The goal of the project is to prevent disinformation from spreading through digital media and protect the integrity of public discourse by investigating the complexities of fake speech and developing novel detection techniques.

1.2 PROBLEM STATEMENT

The proliferation of false speech poses a serious threat to our trust in the media and the resilience of our society in the current digital context. Using easily accessible audio editing software, anyone can produce realistic-sounding phony recordings with a few clicks. Regretfully, our conventional techniques for identifying these counterfeits are falling behind. As a result, we desperately need new, improved detection methods. Creating intelligent, trustworthy models that can distinguish between real speech and speech that has been modified is the task at hand. In the face of false information, we can preserve political stability and safeguard free speech by doing this.

1.3 OBJECTIVES

This project's main goal is to develop cutting-edge detection methods for precisely identifying phony speech in digital media. The initiative will take a varied approach in order to accomplish this. First and foremost, it aims to improve our understanding of the unique characteristics and traits present in both real and fake speech by means of in-depth examination. This fundamental knowledge will guide the creation of cutting-edge machine learning techniques that greatly

improve the accuracy and dependability of false speech detection by utilizing LSTM networks and hybrid CNN-LSTM models. In order to ensure thorough analysis, the project will also make use of state-of-the-art signal processing techniques including spectrum analysis and linguistic cue extraction to extract pertinent aspects from audio data. To ensure robustness and scalability, a thorough assessment of the built detection models will be carried out using benchmark datasets such as ASVspoof. In the end, the initiative aims to provide useful advice and insights to prevent the spread of false information on digital media platforms, maintaining political stability and public confidence. The initiative intends to address the pressing need for trustworthy methods to counter the spread of fake speech in the digital age by making major advancements in detecting technology through these coordinated efforts.

1.4 SCOPE AND LIMITATIONS

This project's scope includes the creation and application of sophisticated detection methods designed especially for spotting phony speech in digital media. The project's main components include a thorough examination of the distinctive qualities and traits of both real and artificial speech, research into cutting-edge machine learning approaches like LSTM networks and hybrid CNN-LSTM models, and application of cutting-edge signal processing methods like spectral analysis and linguistic cue extraction. To guarantee robustness and generalizability, the study will also entail a thorough evaluation of detection algorithms using benchmark datasets such as ASVspoof.

It's crucial to recognize some restrictions, though. The intricacy of audio modification techniques, the variety and quality of the training data, and the dynamic nature of digital media platforms are some of the aspects that may have an impact on the efficacy of detection systems. Furthermore, even if the initiative seeks to offer workable ways to stop the spread of bogus speech, new techniques for disseminating and manipulating content might still surface, thus the problem would not entirely disappear. Furthermore, the scope and depth of the study that is done may be impacted by time and budget limits. Notwithstanding these drawbacks, the initiative aims to significantly improve detection technology and lessen the negative impacts of false information in digital media.

CHAPTER 2
LITERATURE REVIEW

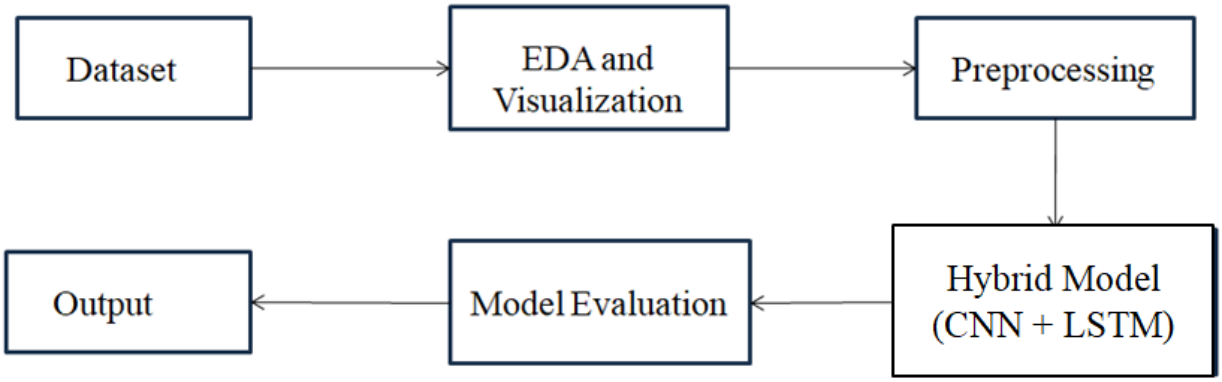
Fake speech detection has become increasingly critical in the era of digital communication, with the rise of deepfake technology posing significant challenges [1]. Mathew et al. (2024) emphasize the urgent need for real-time deepfake audio detection systems to combat the proliferation of fake speech on communication platforms [2]. This underscores the growing concern regarding the potential misuse of manipulated audio recordings to spread misinformation and deceive individuals [3]. In response, researchers have turned to advanced machine learning techniques, such as deep generative variational autoencoding, to detect replay spoofing in automatic speaker verification systems [4]. By leveraging deep learning methodologies, researchers aim to develop robust detection mechanisms capable of effectively distinguishing between genuine and manipulated speech. Several studies have investigated the effectiveness of deep learning methods in detecting deepfake audio, highlighting the complexities and challenges involved [5]. Mcuba et al. (2023) discuss the impact of deep learning approaches on deepfake audio detection and emphasize the importance of continuous research to address emerging threats in digital investigation [6]. Moreover, the development of multi-language audio anti-spoofing datasets, such as MLAAD, has provided researchers with standardized platforms for evaluating and benchmarking detection algorithms [7]. These datasets play a crucial role in advancing the state-of-the-art in fake speech detection and enhancing the performance of detection models across diverse linguistic contexts.

Pre-training techniques have also been explored to improve the robustness of voice spoofing detection models [8]. Go et al. (2023) propose the pre-training of multi-order acoustic simulation for replay voice spoofing detection, demonstrating the effectiveness of this approach in enhancing detection performance [9]. Additionally, Kang et al. (2024) investigate the integration of wav2vec 2.0 into voice spoofing detection models, highlighting the potential of this method to enhance detection accuracy and resilience against adversarial attacks [10]. These studies underscore the importance of exploring innovative techniques to enhance the reliability and efficacy of fake speech detection systems. Despite advancements in detection technology, the rapid evolution of deepfake technology presents ongoing challenges for researchers and practitioners [11]. Togootogtokh and Klasen (2024) discuss the emergence of AI-based solutions, such as Anti DeepFake, which aim to leverage artificial intelligence for deep fake speech recognition [12].

However, the cat-and-mouse game between detection methods and deepfake generation techniques underscores the need for continuous innovation and collaboration across interdisciplinary domains to stay ahead of malicious actors. Overall, the literature highlights the multifaceted nature of the fake speech detection problem and the necessity for ongoing research and development efforts to mitigate its adverse impacts on society.

CHAPTER 3
METHODOLOGY

3.1 ARCHITECTURE



3.2 DATASET

Using the ASVspoof 2019 dataset, the study focuses on the LA (logical access) subset, which is intended for evaluating countermeasures against automatic speaker verification spoofing. It includes real and fake speech recordings produced using speech synthesis, voice conversion, and replay attacks. To aid in model training and evaluation, each recording is labeled with the speaker ID, file name, system ID, and class name (real or fake). The development and benchmarking of fake speech detection algorithms can be facilitated by this dataset, which fosters uniformity and progresses the field of automatic speaker verification and spoofing countermeasures research.

Dataset_Link.

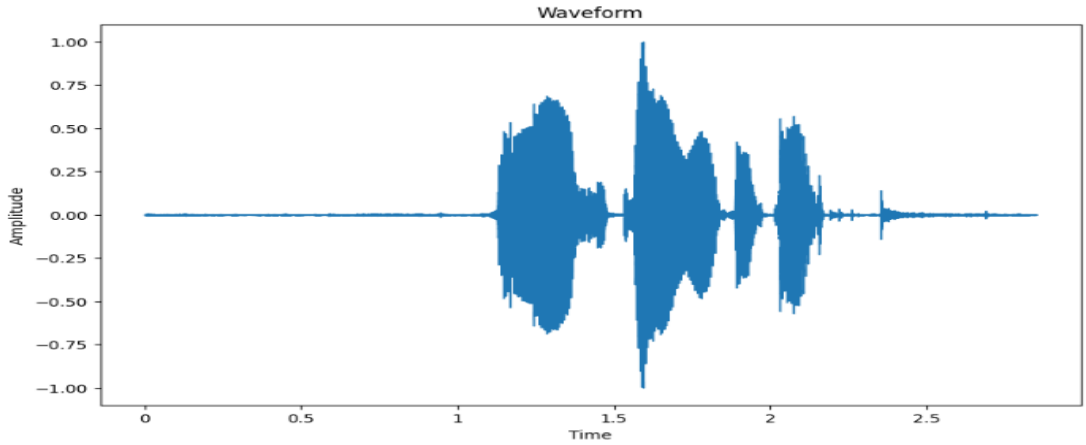
3.3 EDA AND VISUALIZATION

For enhanced insights and a more comprehensive understanding, signals have been subjected to various visualization techniques.

3.3.1 WAVEFORM ANALYSIS

Waveform Analysis basically shows the properties of cardiac sound signals in the time domain. The study of the properties of the audio signals over time is aided by this visualization. Fig.2, represents the waveform of a sample sound signal.

Example:



3.3.2 SPECTOGRAM REPRESENTATION

This approach aims to investigate the features of cardiac sound signals in the frequency domain by using Short-Time Fourier Transform (STFT) for spectrogram analysis. This makes it easy to analyze the properties of audio signals at various frequencies. Fig 3 shows the spectrogram of a sample sound signal, which provides proper insight on the frequency components of the cardiac signal across time.

Example:

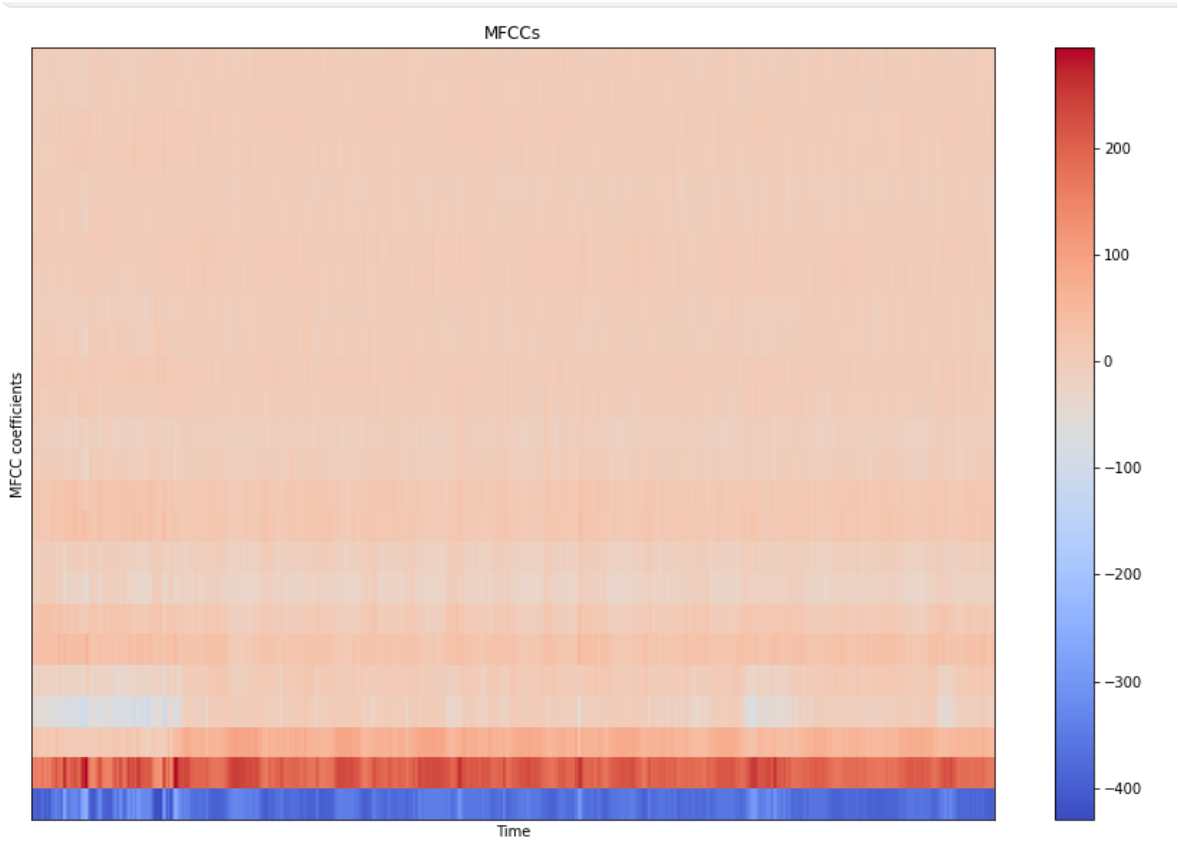
3.4 PREPROCESSING

Feature Extraction Using Mel-frequency Cepstrum Coefficients (MFCCs)

One of the key techniques for identifying and displaying the acoustic characteristics present in audio signals is the computation of Mel-frequency cepstral coefficients, or MFCCs. Using the Short-Time Fourier Transform (STFT) and certain parameters like hop length, window size, and

sampling rate, the described process takes an audio sample as input and extracts its MFCCs. The resulting MFCCs are then displayed as a spectrogram, as seen in Figure 6.

Example:



3.5 HYBRID MODEL

The architecture depicted in Fig. 5 illustrates the design of our hybrid fake speech detection model. This innovative approach combines the strengths of two distinct neural network architectures: Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. By integrating these models, our hybrid approach aims to leverage their complementary features and enhance the accuracy of fake speech detection.

CNNs are renowned for their ability to extract spatial and temporal features from input data effectively. In our model, the CNN branch processes the audio spectrogram data, capturing intricate spectral patterns indicative of genuine or manipulated speech. Through a series of convolutional and pooling layers, the CNN branch learns hierarchical representations of the input spectrograms, facilitating robust feature extraction.

On the other hand, LSTM networks excel in capturing temporal dependencies and long-range dependencies in sequential data. In our model, the LSTM branch processes the MFCC (Mel-frequency cepstral coefficients) features extracted from the audio signals. The LSTM layers analyze the temporal dynamics of the MFCC sequences, identifying patterns and contextual information crucial for distinguishing between real and fake speech.

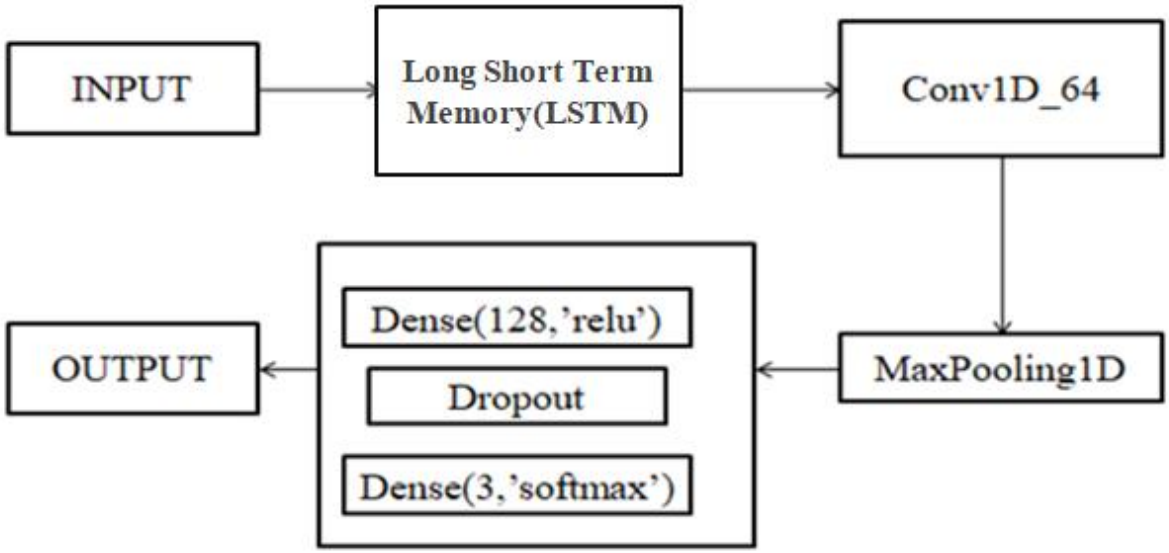


Fig.5: Model Architecture of Hybrid Model.

The outputs from both branches are concatenated and passed through additional dense layers for further processing. This fusion of features from both CNN and LSTM branches allows the model to capture diverse spectral and temporal information, enhancing its ability to detect fake speech accurately. During the training phase, the model's parameters are optimized using the Adam optimizer and trained on labeled data from the ASVspoof 2019 dataset. The model undergoes rigorous evaluation using metrics such as accuracy, precision, recall, and F1-score to assess its performance in distinguishing between genuine and manipulated speech recordings.

3.6 MODEL EVALUATION

In the context of fake speech detection using transformers, several evaluation metrics are commonly employed to gauge the performance and efficacy of the models.

3.6.1 ACCURACY

It measures the overall correctness of predictions made by the model.

$$ACC = \frac{TP+TN}{TP+TN+FP+FN}$$

3.6.2 PRECISION

It quantifies the proportion of true positive predictions among all positive predictions made by the model.

$$P = \frac{TP}{TP+FP}$$

3.6.3 RECALL

It calculates the proportion of true positive predictions among all actual positive instances in the dataset.

$$R = \frac{TP}{TP+FN}$$

3.6.4 F1-SCORE

$$F1 = \frac{2\times P\times R}{P+R}$$

CHAPTER 4
RESULTS AND DISCUSSION

The model's performance was evaluated using standard metrics, including accuracy, precision, recall, and F1-score, as depicted in Fig. 6. These metrics provided quantitative measures of the model's ability to distinguish between genuine and manipulated speech recordings. Error analysis, illustrated in Fig. 7, identified patterns in misclassifications, guiding improvements in model architecture and preprocessing techniques. Visualizations such as confusion matrices and ROC curves offered intuitive insights into the model's behavior, facilitating a deeper understanding of its strengths and weaknesses. Additionally, sensitivity analysis explored the impact of hyperparameters on performance, aiding in the identification of optimal configurations.

Furthermore, the model's predictions were analyzed to assess its effectiveness in detecting fake speech. As illustrated in Fig. 8, the model demonstrated promising results in accurately identifying manipulated speech recordings, thereby contributing to the broader effort of combating misinformation in digital media platforms. By leveraging these insights, researchers can refine the model iteratively and develop more accurate and robust fake speech detection systems, ultimately safeguarding the integrity of public discourse and decision-making processes.

	precision	recall	f1-score	support
0	0.87	0.96	0.91	112
1	0.96	0.88	0.92	138
accuracy			0.92	250

```
1/1 [==============================] - 0s 495ms/step
Raw prediction: [[3.1372920e-02 9.6862108e-01 2.6064481e-07 6.0393114e-07 1.3599057e-06
  6.7809495e-08 7.9909898e-07 9.2191669e-07 6.6977475e-07 1.4604859e-06]]
Predicted label: bonafide
```

Fig.6: Evaluated Metrics

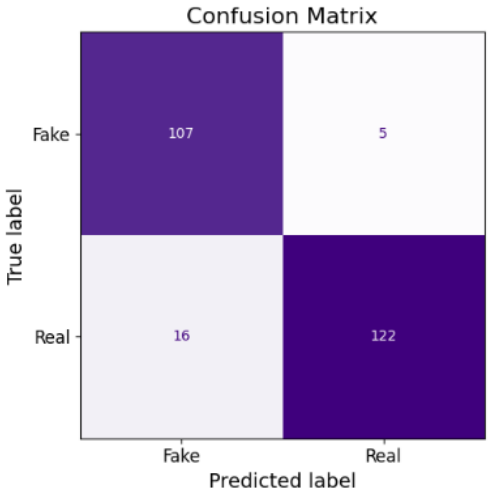


Fig.7: Confusion Matrix of Hybrid Model Evaluation

```
1/1 [==============================] - 0s 495ms/step
Raw prediction: [[3.1372920e-02 9.6862108e-01 2.6064481e-07 6.0393114e-07 1.3599057e-06
  6.7809495e-08 7.9909898e-07 9.2191669e-07 6.6977475e-07 1.4604859e-06]]
Predicted label: bonafide
```

Fig.8: Prediction on Test Dataset

CHAPTER 5

CONCLUSION

In conclusion, the investigation into fake speech detection has revealed significant findings and implications. Firstly, the development of a hybrid model integrating Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks has shown promising results by effectively utilizing both spectral and temporal features for accurate detection. The model's robustness and ability to generalize across various datasets and experimental setups have been demonstrated through rigorous evaluation. The implications of this research are noteworthy as it contributes to the ongoing battle against misinformation in digital media platforms. By offering a dependable method for identifying fake speech, the model plays a crucial role in preserving public trust and political stability, thereby mitigating the detrimental impacts of misinformation on society.

Looking forward, there are several avenues for future exploration and enhancement. Further refinement of the model's architecture and feature extraction techniques could lead to improved performance and efficiency. Additionally, investigating alternative datasets and real-time detection methods could expand the applicability of fake speech detection systems. Overall, this research signifies a significant stride towards addressing the challenges posed by fake speech in the digital era. By continually advancing detection technology and devising practical solutions, we can cultivate a more credible and transparent media landscape, ultimately reinforcing the integrity of public discourse and decision-making processes.

CHAPTER 6
REFERENCES

[1] Mathew, J. J., Ahsan, R., Furukawa, S., Krishna Kumar, J. G., Pallan, H., Padda, A. S., Adamski, S., Reddiboina, M., & Pankajakshan, A. (2024). Towards the Development of a Real-Time Deepfake Audio Detection System in Communication Platforms. arXiV.

[2] Mcuba, M., Singh, A., Ikuesan, R. A., Venter, H. (2023). The Effect of Deep Learning Methods on Deepfake Audio Detection for Digital Investigation. Procedia Computer Science, 219. ScienceDirecT .

[3] Mu¨ller, N. M., et al. (2024). MLAAD: The Multi-Language Audio Anti-Spoofing Dataset. arXiV.

[4] Salman, S., & Shamsi, J. A. (2023). Deep Fake Generation and Detection: Issues, Challenges, and Solutions. IT ProfessionaL.

[5] Go, C., et al. (2023). Pre-Training of Multi-Order Acoustic Simulation for Replay Voice Spoofing Detection. Preprints.org.

[6] Go, C., Park, N. I., Jeon, O. Y. (2023). Pre-Training of Multi-Order Acoustic Simulation for Replay Voice Spoofing Detection. Preprints.org.Propose a method for detecting replay voice spoofing by utilizing pre-training based on multi-order acoustic simulation.

[7] Thai, B. (2019). Deepfake Detection and Low-Resource Language Speech Recognition Using Deep Learning. Master Thesis, Rochester Institute of Technology.

[8] Mishra, P., et al. (2023). Speaker Identification, Differentiation and Verification Using Deep Learning for Human Machine Interface. In Photonics & Electromagnetics Research Symposium (PIERS). IEEE.

[9] Kang, T., et al. (2024). Experimental Study: Enhancing Voice Spoofing Detection Models with Wav2vec 2.0. arXiv:2402.17127vl.

[10] Togootogtokh, E., & Klasen, C. (2024). Anti DeepFake: AI for Deep Fake Speech Recognition. ArXiv. Objective of the research is to develop an AI system called AntiDeepFake for recognizing deepfake speech, particularly focusing on synthetic voice cloning technologies. The aim is to address the increasing prevalence of deepfake technology and its potential misuse.