A Novel Approach to Audio Deep Fake Detection using CNN and Bi-LSTM

[1]Vivek Shinde, [2]Rohan Jagtap, [3]Rohan Sonawane, [4]Suyash Yeolekar, [5]Ratna Patil

[1,2,3,4,5] Department of Artificial Intelligence and Data Science, Vishwakarma Institute of Information Technology,
Pune, Maharashtra, India.

Email: [1]shinde.22110036@viit.ac.in, [2]randhir.22111304@viit.ac.in, [3]rohan.22111303@viit.ac.in, [4] suyash.22111297@viit.ac.in,
[5]ratna.patil@viit.ac.in

Abstract: The development of deepfake technology Shows a serious threat to the integrity of voice communications. with potential misuse in identity theft, misinformation, and fraud. The main goal of the project is to create a powerful detection system that can distinguish real sound using the audio signal to increase the security of audio files. To achieve this, we propose a novel approach that integrates Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (Bi-LSTM) networks. The CNN component extracts spatial features from audio spectrograms, identifying unique patterns that differentiate real and fake audio. These extracted features are then processed by the Bi-LSTM, which captures temporal dependencies and context, further improving detection accuracy. The proposed model was tested on the “Lj speech dataset” and “wavefake dataset” , demonstrating superior performance over conventional methods with improved generalization across various spoofing techniques. This approach is particularly applicable in enhancing security in voice authentication systems, digital forensics, and other domains requiring reliable detection of audio manipulations.

Keywords: Audio Deep Fake Detection, CNN, Bi-LSTM, Audio Spectrograms, Voice Authentication Security, Digital Forensics, Spoofing Techniques Detection.

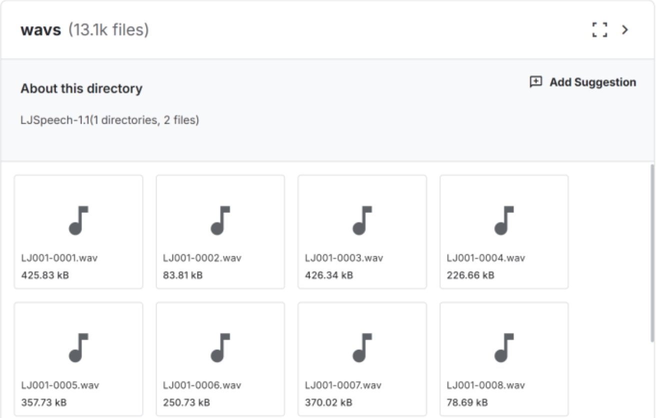
I. INTRODUCTION

The exponential growth in deep learning capabilities has facilitated the production of exceptionally lifelike synthetic audio, complicating the task of differentiating between artificially generated and authentic vocal content. This technological leap presents opportunities for various sectors, including entertainment and accessibility, despite the increasing complexity in identifying genuine speech from fabricated alternatives. it also raises significant concerns regarding information authenticity, privacy, and security. Deepfake audio has the potential to be misused in harmful ways, such as identity theft, fraud, and misinformation, which highlights the pressing need for accurate and reliable detection mechanisms. Traditional audio verification methods, which often rely on basic acoustic feature comparisons, are proving inadequate against sophisticated deepfake algorithms.

We introduced a deep learning method to solve these problems that integrates CNN with BiLSTM networks to efficiently detect deepfake audio. Our system employs a comprehensive array of acoustic features, including MFCC, Mel spectrograms, Constant Q Cepstral Coefficients (CQCC), and Constant-Q Transform (CQT) vectors. This diverse feature set enables the capture of both spatial and temporal aspects of audio signals. The CNN layers extract spatial patterns, while the BiLSTM layers capture temporal dependencies, creating a comprehensive representation of audio characteristics. Evaluated on benchmark datasets like ASVSpoof 2019 and FoR, the model demonstrates superior detection performance, as measured by accuracy and Equal Error Rate (EER). By leveraging both spatial and sequential audio features, this architecture significantly advances the reliability of audio deepfake detection, supporting efforts to preserve digital audio authenticity.

II. DATASET

For this research, to guarantee a strong and varied training and testing environment, we used a variety of available datasets for our study. The latter included the WaveFake-Test dataset as well as the LJSpeech-1.1 dataset.



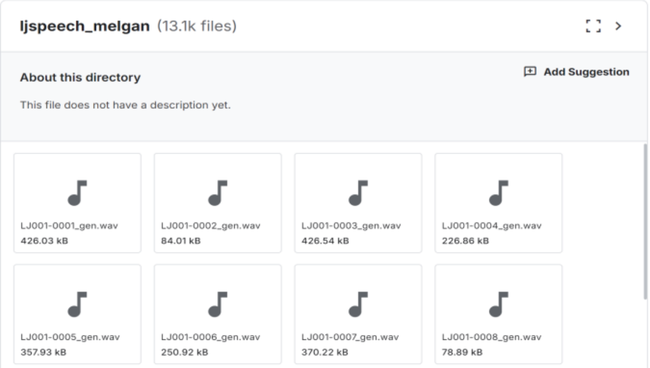
wavs (13.1k files)

About this directory

LJSpeech-1.1(1 directories, 2 files)

LJ001-0001.wav 425.83 kB
LJ001-0002.wav 83.81 kB
LJ001-0003.wav 426.34 kB
LJ001-0004.wav 226.66 kB
LJ001-0005.wav 357.73 kB
LJ001-0006.wav 250.73 kB
LJ001-0007.wav 370.02 kB
LJ001-0008.wav 78.69 kB

Fig. 1: LJSpeech-1.1 dataset [1]



ljspeech_melgan (13.1k files)

About this directory

This file does not have a description yet.

LJ001-0001_gen.wav 426.03 kB
LJ001-0002_gen.wav 84.01 kB
LJ001-0003_gen.wav 426.54 kB
LJ001-0004_gen.wav 226.86 kB
LJ001-0005_gen.wav 357.93 kB
LJ001-0006_gen.wav 250.92 kB
LJ001-0007_gen.wav 370.22 kB
LJ001-0008_gen.wav 78.89 kB

Fig. 2: WaveFake-Test dataset [2]

The LJSpeech-1.1 Corpus contains a rather large number of high quality audio recordings of an individual reading sections of various books. It provides real life audio examples. Indeed, the WaveFake-Test dataset has sampling audio of deepfake sourced from advanced technologies such as MelGAN synthesis. We managed to merge these two datasets and create the final dataset that contained synthetic and real audio, which helped the model to be trained against both the audio types.

III. RELATED WORK

Many papers have contributed to the development of new techniques for deep sound detection using CNNs and Bi-LSTMs, this time using different methods and techniques for solving local problems. Here we collect many important information in this field.

1) (Taiba Majid Wani et al.,2024) [3]: An innovative approach for identifying deepfakes through the integration of CNN and BiLSTM technologies. Our method uses various acoustic features including MFCC, Mel spectrograms, CQCC, and CQT processed by CNN and analyzed by BiLSTM to detect sugar and body structure. Our high quality and low error rate (EER) are validated on ASVSpoof 2019 and FoR datasets, improving the ability to capture deepfakes.

2) (Ousama A. Shaaban [4]: In the first half of this study, where its use in creating and detecting audio deepfakes is examined, information about generic deepfakes is provided. The second section elucidates and contrasts the primary methods for creating audio deepfakes. This investigation's findings encompass diverse approaches, including ML and DL algorithms for deepfake detection, social media analysis, and numerical property evaluation. The principal techniques for identifying fake news in these studies comprise SVM, decision tree, CNN, Siamese, power absorption neural networks DNN, and RNN. The effectiveness of these methodologies varies.

3) (Lam Pham et al., 2024) [5]: Developed as part of the EUCINF (European Network and Knowledge) project, a multi-partner European collaboration, we propose a deep learning method to identify deepfakes. We transform the noise signal into a spectrogram using Wavelet Transform (WT), Constant Q Transform (CQT) and Short Time Fourier Transform (STFT) with various audio filters (Mel, Gammatone, Linear, DCT). We use (2) Multilayer Perceptron (MLP) to analyze audio embeddings from pre-learned models (Whisper, Seamless, Speechbrain, Pyannote); (3) Transfer learning using visual models (e.g. ResNet-18, MobileNet-V3, EfficientNet); and (4) Base CNN, RNN and C-RNN models. The combination of the best model from the ASVspoof 2019 dataset achieved an EER of 0.03, demonstrating the effectiveness of spectrogram modification and deep learning on discovered audio deepfakes.

4) (Lian Huang et al.,2020) [6]: The automatic speaker verification (ASV) system is not very effective in detecting replay attempts and is quite vulnerable to spoofing attacks. We propose an attention-augmented DenseNet-BiLSTM model that uses segmentation-based linear filter bank features to address this problem. Using a zero-crossing rate and short-term energy, our method first finds the silent segments of the speech signal. If there is not enough silent data, decaying tails are chosen. These segments are then used to extract characteristics of a high-frequency linear filter bank. The BTAS 2016 and ASVspoof 2017 datasets were used to test the DenseNet-BiLSTM architecture, which has been improved with attention techniques to reduce overfitting. Relative performance increases of 91.68% and 74.04% on the corresponding datasets demonstrate a notable improvement over baseline systems.

5) (Zaynab M. Almutairi et al.,2023) [7]: Audio deepfakes, in which AI is used to clone voices, pose a significant threat to public safety. Although ML and DL techniques exist for detection, they often require large datasets and extensive preprocessing. To handling these issues, we propose Arabic-AD, a self-supervised learning-based method for detecting synthetic and mimicked Arabic speech. This work presents the first synthetic dataset of a fluent speaker of Modern Standard Arabic (MSA) and includes Arabic recordings from non-native speakers for reliability assessment. In experiments, Arabic-AD achieved outstanding performance with an error rate (EER) of 0.027% and an accuracy of 97%, outperforming current standards without requiring too much data.

6) (Nicholas Wilkinson et al., 2021) [8]: This paper presents a hybrid architecture for speech activity detection (VAD) that combines CNN and BiLSTM layers, trained end-to-end to improve performance in noisy and resource-constrained environments. We use nested k-fold cross-validation to explore hyperparameters and balance model size with performance. Our system, tested on the AVA-Speech dataset, outperforms three baselines, including a larger ResNet model, with an AUC of 0.951, especially in challenging noise conditions. BiLSTM layers provide a ≈2% accuracy improvement over unidirectional LSTM layers.

7) (Ramesh K Bhukya et al.,2024) [9]: the detection of audio deepfakes, which pose significant risks to public safety. Using feature extraction techniques like MFCCs, chromagrams, and Mel-spectrograms, we train ML And DL models on the ASVspoof2021 dataset to classify speech as bonafide or spoofed. Experimental results show the multilayer perceptron achieves 96.07% accuracy, while SVM, k-NN, XGBoost, and RF yield accuracies between 93.54%

and 94.58%. Deep learning models, particularly CNN, outperform others with an accuracy of 98.2%, highlighting its effectiveness in detecting deepfake audio.

8) (Rami Mubarak et al.,2023) [10]: This study explores the growing threat of deepfakes (fake image, audio, and text content) generated using AI techniques. While the focus is on image and audio deepfakes, text deepfakes are also becoming a major concern due to advances in NLP.The study examines the social, political, economic, and technological impacts of deepfakes and critically evaluates current detection methods. It calls for unified, real-time solutions and emphasizes the need for a comprehensive approach that combines technical measures, public awareness, and legislative action to address the challenges of deepfakes.

9) (Mouna Rabhi et al., 2024) [11]: This paper examines the Weakness of AI-based audio authentication systems to adversarial attacks using deepfake audio. We demonstrate that state-of-the-art audio deepfake classifiers, such as the Deep4SNet model with 98.5% accuracy, can be severely compromised. Our two adversarial attacks, exploiting generative adversarial networks, reduce the accuracy to nearly 0%, especially to 0.08% in a gray-box scenario. To counter these attacks, we propose a lightweight and generalizable defense mechanism that can be adapted to any deepfake audio detector and outline future research directions Improves security for audio-based authentication systems.

This project builds on existing research by proposing a novel hybrid CNN-Bi-LSTM architecture that focuses on performance and attack for real-world noise optimization. Through the combination of CNN and Bi-LSTM techniques, the method attempts to detect difference and proximity, improve detection value in the noise environment, and prevent adversarial interference.

IV. METHODOLOGY

1] Preprocessing of Audio Data

To analyze deep sound, pre-processing is essential to remove important elements of the original sound. In this study, MFCC, a subtraction method known for its ability to imitate the human ear's response to sound, was used. The detailed steps for extracting MFCCs are as follows:

1.1. Audio Framing and Windowing :

The original audio waveform is divided into overlapping lines of short duration (usually 20-40 milliseconds). A Hamming window w(n) is applied to each frame to reduce spectral leakage at the boundary:

$$w(n) = 0.54 - 0.46\, cos\left(\frac{2\pi n}{N-1}\right) \qquad (1)$$

where NNN is the number of samples in the frame.

1.2. Short-Time Fourier Transform (STFT):

Each frame window is passed through a STFT, which converts the signal recorded over time into a frequency:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi kn/N} \qquad (2)$$

where x(n) is the time-domain signal and k represents the frequency bin. The power spectrum |X(k)|^2 is computed for each frame.

1.3. Mel Filter Bank:

The frequency domain representation is mapped to the Mel scale using a series of triangular filter banks. The Mel scale approximates human auditory perception, with a greater emphasis on lower frequencies. The conversion of the frequency f to Hertz on the Mel scale m(f) is given by:

$$m(f) = 2595 log_{10}\left(1 + \frac{f}{100}\right) \qquad (3)$$

The power spectrum is passed through these filters, and the energy in each filter EmE_mEm is calculated.

1.4. Logarithm of Filter Bank Energies:

To compress the range of the filter bank energies, a logarithmic function is applied:

$$LogEnergy(m) = log(E_m + \epsilon) \qquad (4)$$

where ϵ\epsilon is a small constant to prevent taking the logarithm of zero.

1.5. Discrete Cosine Transform (DCT):

The final step in extracting MFCCs is applying the DCT to the log energies:

$$MFCC(c) = \sum_{m=0}^{M-1} LogEnergy(m) cos\left[\frac{\pi c}{M}(m + 0.5)\right] \qquad (5)$$

where c is the cepstral coefficient index. Typically, the first 12-13 coefficients are retained, representing the most significant audio features.

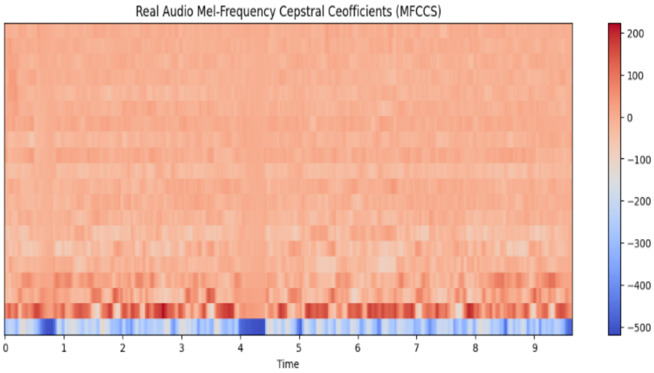


Fig. 3: Real Audio MFCCS

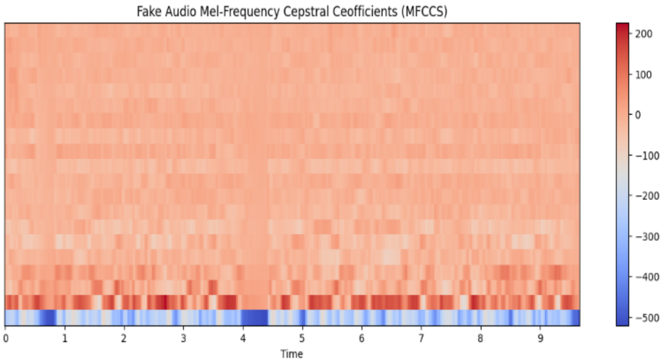
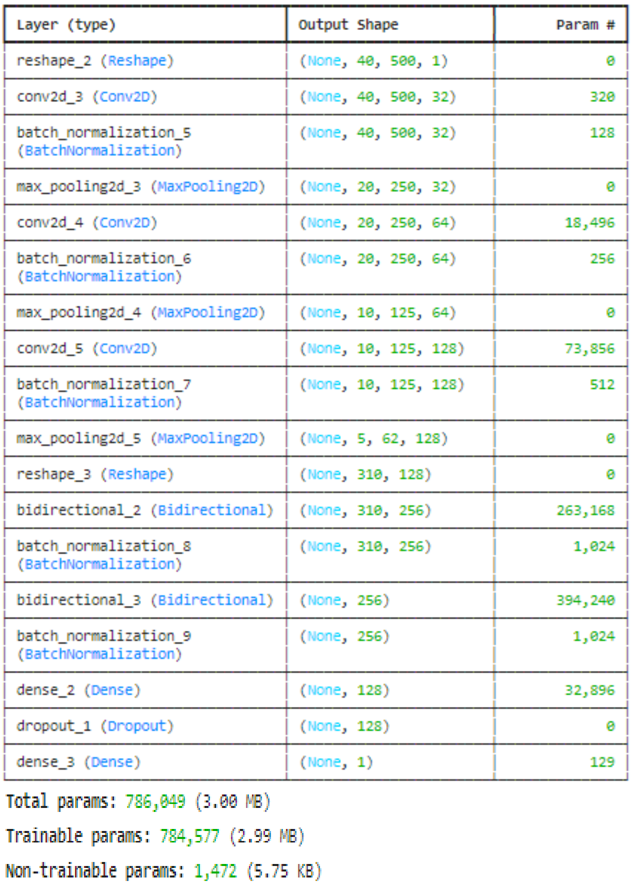


Fig. 4: Fake Audio MFCCS

2] Model Architecture

The proposed model architecture for deep audio forgery detection combines a CNN for feature extraction and BiLSTM layers for sequence modeling. This combined approach exploits the spatial feature learning capabilities of CNN and the temporal sequence modeling power of BiLSTM.



Layer (type)	Output Shape	Param #
reshape_2 (Reshape)	(None, 40, 500, 1)	0
conv2d_3 (Conv2D)	(None, 40, 500, 32)	320
batch_normalization_5 (BatchNormalization)	(None, 40, 500, 32)	128
max_pooling2d_3 (MaxPooling2D)	(None, 20, 250, 32)	0
conv2d_4 (Conv2D)	(None, 20, 250, 64)	18,496
batch_normalization_6 (BatchNormalization)	(None, 20, 250, 64)	256
max_pooling2d_4 (MaxPooling2D)	(None, 10, 125, 64)	0
conv2d_5 (Conv2D)	(None, 10, 125, 128)	73,856
batch_normalization_7 (BatchNormalization)	(None, 10, 125, 128)	512
max_pooling2d_5 (MaxPooling2D)	(None, 5, 62, 128)	0
reshape_3 (Reshape)	(None, 310, 128)	0
bidirectional_2 (Bidirectional)	(None, 310, 256)	263,168
batch_normalization_8 (BatchNormalization)	(None, 310, 256)	1,024
bidirectional_3 (Bidirectional)	(None, 256)	394,240
batch_normalization_9 (BatchNormalization)	(None, 256)	1,024
dense_2 (Dense)	(None, 128)	32,896
dropout_1 (Dropout)	(None, 128)	0
dense_3 (Dense)	(None, 1)	129

```
Total params: 786,049 (3.00 MB)
Trainable params: 784,577 (2.99 MB)
Non-trainable params: 1,472 (5.75 KB)
```

Fig. 5: Model Architecture

2.1. CNN Layers for Feature Extraction:

The model starts with a series of convolutional and pooling layers:

- Conv2D layer: The first set of CNN layers extracts hierarchical features from the input MFCCs. The initial layer consists of 32 filters with kernel size (3, 3), followed by ReLU activation and MaxPooling with pool size (2, 2) to sample the feature maps
- Following the Conv2D layers: Two additional convolutional layers with 64 and 128 filters respectively continue to extract features, each followed by ReLU activation and max pooling.

2.2. Normalization and Regularization:

Batch normalization is applied after each iteration to improve model detail and training stability. Dropout layers are used to prevent overfitting.

2.3. Temporal Modeling with BiLSTM Layers:

While maintaining the time-step dimension, the CNN layers' output is transformed into a 3D tensor appropriate for LSTM input. The following is how the BiLSTM layers are set up:

- With 64 units in each direction, the first BiLSTM layer returns sequences to record dependencies across time.
- The second BiLSTM layer summarizes sequence information into a single output by processing the first BiLSTM's output.

2.4. Layers with Complete Connectivity:

The final output from the BiLSTM is processed by:

- Dense Layer: To penalize big weights, this fully linked layer has 128 units, ReLU activation, and L2 regularization.
- Output Layer: A dense layer that performs binary classification (actual vs. false audio) using a sigmoid activation function.

Combining CNN and BiLSTM layers allows the model to benefit from the spatial feature extraction of CNNs and the sequential modeling of BiLSTMs. This hybrid approach is particularly useful in audio deepfake detection, where:

- CNNs identify short-term spectral patterns (e.g., characteristic artifacts in the frequency domain)
- BiLSTMs analyze temporal dependencies and patterns across the audio frames, enabling the model to understand the progression of audio features over time.

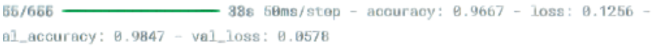
This integration helps address the challenge of detecting subtle temporal inconsistencies in deep fake audio that might not be captured by CNN or LSTM alone.

3. Model Training:

The model was compiled using the Adam optimizer That has learning rate of 0.001, a binary cross-entropy loss, and an accuracy measure of . Training was performed with early recall and learning rate reduction to optimize convergence .

V. RESULTS

Within the scope of the research, we developed for deep voice recognition CNN-BiLSTM model, which accepts MFCC as input, and trained it on LJ Speech and WaveFake datasets. The model has shown good performance on test data completing with a final accuracy of 98.4% as illustrated in Fig. 6 (see training list). The loss for this model is however changed to 0.0578 meaning that the model appears to be well adjusted for the training data yet resilient to overfitting.



```
55/666 ━━━━━━━━━━━━━━━━━━━━ 33s 50ms/step - accuracy: 0.9667 - loss: 0.1256 -
al_accuracy: 0.9847 - val_loss: 0.0578
```

Fig.6 Validation Accuracy

We applied early stopping whenever the validation accuracy plateaued for 10 epochs, which was essential for avoiding overfitting and achieving efficient convergence.

Fig. 7 displays the training and validation accuracy trends over 45 epochs. The model's validation accuracy rapidly increased during the initial epochs, stabilizing above 95% with minor fluctuations. Notably, the validation accuracy remained consistent with the training accuracy, suggesting effective generalization and low overfitting. The CNN component effectively captures the spatial features of the MFCC input, while the BiLSTM layer exploits temporal patterns, improving the model's sensitivity to subtle differences between real and fake sounds.

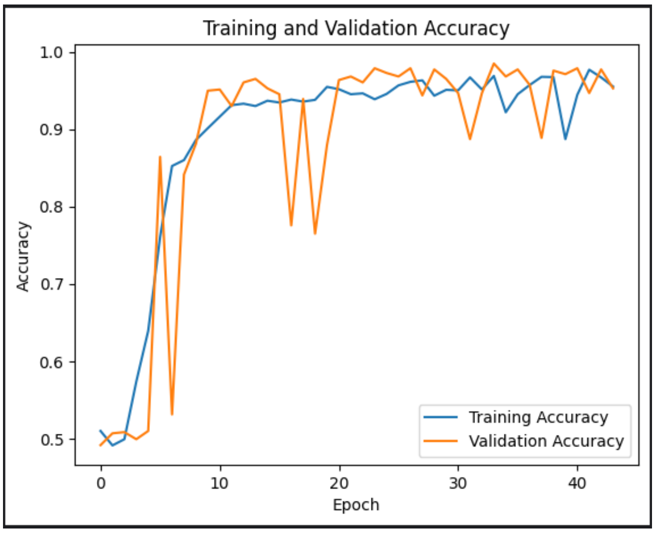


Fig. 7 Training and Validation Accuracy graph

The findings affirm that our CNN-BiLSTM model is a viable option for fake audio detection and is a practical solution in deepfake audio content detection.

VI. DISCUSSION

The goal of this paper is to improve detection of real and manipulated audio through incorporating two deep learning approaches: Bi-LSTM networks and CNNs. Milestones achieved during this work include modifications to the separation of target audio from the background, leading to more effective detection. Audio is represented in this architecture as an image, then CNN's structural approach is integrated into the Bi-LSTM network focused on time sequences. This approach subsequently enables us to decipher the two-dimensional patterns of interest embedded within audio signals. The combination of aforementioned factors strengthens the model due to ease of detecting small differences that traditional detection methods fail to detect (Bi-LSTM) and spatial feature extraction (CNN).

The system allows an effective deepfake clip scanning and reduces the huge computable requirement since this process is made asymmetric and hence is friendly for large scale applications. Our work demonstrates how hybrid models can be used instead of deep learning techniques to enahance the accuracy and reliability of deep learning systems**.** Our work shows how the accuracy and reliability of deepfake detection systems can be enhanced through the application of the hybrid model amidst the changing tactics of deepfakes. It demonstrates how critical it is to represent concepts and suitable models of the networks capturing the spaces of audio features.

These findings of this experiment are a concern for future applications that are meant to help in fighting fraud and in the field of justice and security. Also, our research participates in a much wider debate on blocking electronic devices by demonstrating deep learning techniques. Future releases may take advantage of hearing more voices, using more data, and using common strategies that offer the model other methods of research to augment its generality. Finally, our work provides a foundation for sequential steps to be taken in this important field, by establishing the ability of novel machine learning techniques to protect qualities associated with voice transmissions.

CONCLUSION

We propose an innovative approach in this study that employs deep learning to detect altered audio and be able to tell between them, the altered and the real. The novel approach involves the integration of a CNN and a Bi-LSTM network, allowing the audio data to exploit both spatial features and temporal elements as well. This makes it easier than conventional methods that normally make use of signal processing or other machine learning techniques, to conduct a verification process. We address the principal gap concerning the identification in the scanning strategy by introducing MFCCs as input, which produces a combination that enhances effectiveness. Our results durability in practical situations as well as showcase great enhancement in the sensing non-monotonic sounds which might differ from normal speech and the usage of many humans. Despite the good performance and ability of excellent performance there are shortcomings including insufficient data that will be augmented to meet up with the breadth of the development process and challenges of scalability for on-the-spot performance. Nevertheless, all these gaps are addressed by our work, and it creates possibilities for further research, such as studying alternative approaches (e.g. spectrograms), utilizing learning models, and speeding up the process for more extensive application usage.

REFERENCES

[1] https://www.kaggle.com/datasets/mathurinache/the-lj-speech-dataset

[2] https://www.kaggle.com/datasets/andreadiubaldo/wavefake-test

[3] T. M. Wani, S. A. A. Qadri, D. Comminiello, and I. Amerini, "Detecting Audio Deepfakes: Integrating CNN and BiLSTM with Multi-Feature Concatenation," *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec '24)*, Association for Computing Machinery, New York, NY, USA, pp. 271–276, 2024. doi: 10.1145/3658664.3659647.

[4] O. A. Shaaban, R. Yildirim and A. A. Alguttar, "Audio Deepfake Approaches," in *IEEE Access*, vol. 11, pp. 132652-132682, 2023, doi: 10.1109/ACCESS.2023.3333866.

[5] L. Pham, P. Lam, T. Nguyen, H. Nguyen and A. Schindler, "Deepfake Audio Detection Using Spectrogram-based Feature and Ensemble of Deep Learning Models," *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*, Erlangen, Germany, 2024, pp. 1-5, doi: 10.1109/IS262782.2024.10704095.

[6] L. Huang and C. -M. Pun, "Audio Replay Spoof Attack Detection by Joint Segment-Based Linear Filter Bank Feature Extraction and Attention-Enhanced DenseNet-BiLSTM Network," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1813-1825, 2020, doi: 10.1109/TASLP.2020.2998870.

[7] Z. M. Almutairi and H. Elgibreen, "Detecting Fake Audio of Arabic Speakers Using Self-Supervised Deep Learning," in *IEEE Access*, vol. 11, pp. 72134-72147, 2023, doi: 10.1109/ACCESS.2023.3286864.

[8] N. Wilkinson and T. Niesler, "A Hybrid CNN-BiLSTM Voice Activity Detector," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, 2021, pp. 6803-6807, doi: 10.1109/ICASSP39728.2021.9415081.

[9] R. K. Bhukya, A. Raj and D. N. Raja, "Audio Deepfakes: Feature Extraction and Model Evaluation for Detection," *2024 5th International Conference for Emerging Technology (INCET)*, Belgaum, India, 2024, pp. 1-6, doi: 10.1109/INCET61516.2024.10593405.

[10] R. Mubarak, T. Alsboui, O. Alshaikh, I. Inuwa-Dutse, S. Khan and S. Parkinson, "A Survey on the Detection and Impacts of Deepfakes in Visual, Audio, and Textual Formats," in *IEEE Access*, vol. 11, pp. 144497-144529, 2023, doi: 10.1109/ACCESS.2023.3344653

[11] M. Rabhi, S. Bakiras, and R. Di Pietro, "Audio-deepfake detection: Adversarial attacks and countermeasures," *Expert Systems with Applications*, vol. 250, 2024, Art. no. 123941. [Online]. Available: https://doi.org/10.1016/j.eswa.2024.123941

★★★