Deepfake Voice Detection Using Machine Learning

Amruthesh S G UG Student, Department of Information Science and Engineering, Sir M Visvesvaraya Institute of Technology, Bengaluru, Karnataka, India
Maninder Kaur UG Student, Department of Information Science and Engineering, Sir M Visvesvaraya Institute of Technology, Bengaluru, Karnataka, India
Tejaswini G H UG Student, Department of Information Science and Engineering, Sir M Visvesvaraya Institute of Technology, Bengaluru, Karnataka, India
Mokshaprada P UG Student, Department of Information Science and Engineering, Sir M Visvesvaraya Institute of Technology, Bengaluru, Karnataka, India
Ms.Sowjanya Lakshmi. A Assistant Professor, Department of Information Science and Engineering, Sir M Visvesvaraya Institute of Technology, Bengaluru, Karnataka, India
Dr G C Bhanu Prakash Professor and Head of Department, Department of Information Science and Engineering, Sir M Visvesvaraya Institute of Technology, Bengaluru, Karnataka, India

Abstract

With the advent of deep learning, audio synthesis techniques has advanced so quickly that it's hard to tell real speech from fake ones. In this work, we leverage voice-conversion and synthesis approaches to design and implement a robust method for fake speech detection targeted towards LA threats. To increase size of dataset and generalize model, the system uses deep learning model with data augmentation that includes techniques like time stretching, pitch shifting and volume scaling. Some of the pre-processing steps include normalizing, removing noise, and converting the audio (segmented in chunks of 4 seconds) into ABRs.Melspectrograms obtained by Znormalization and the Fast Fourier Transform (FFT) are used as feature representations. The proposed architecture consists of multiple layers of convolutions, both 2D and 1x1 convolutions, batch normalization, max-pooling, ReLU activation and fully connected layers. Dropout and other possible sources of regularisation are used to increase the model resistance to overfitting. We used ASVspoof 2019 corpus for the training and testing of the model and further add the various techniques to the simulate the real world. For the analysis of classification, confusion matrix along with accuracy, precision, recall, F1-score and ROC-AUC were used. Results indicate the system was effective in discriminating between genuine and deceptive speech, and had high levels of accuracy in detecting deception.Our goal is to create an artificial speech recognition system that uses deep learning to differentiate between real and fake speech. Our method makes advantage of the ASVspoof 2019 benchmark, which comprises an extensive set of spoof and real speech samples. We use Mel-Frequency Cepstral Coefficients (MFCCs) and spectrograms to extract key components that aid in the detection of significant speech patterns. The learnt characteristics are then examined using a Convolutional Neural Network (CNN), a potent deep learning framework that can identify patterns in both audio and visual data. Unprocessed audio input is converted into MEL spectrograms, preprocessed to reduce noise and improve clarity, and then supplied into the CNN model for classification as part of the training pipeline's systematic process.By providing a scalable, useful, and widely applicable defence against emerging audio spoofing attacks, this study significantly enhances voice-based security systems.

INTRODUCTION

This report provides a comprehensive explanation of the proposed deepfake voice detection system. It begins with the dataset selection, specifically highlighting the use of the ASVspoof 2019 dataset, which offers a wide range of real and spoofed speech samples. Next, the preprocessing methods are detailed, including the conversion of raw audio into MEL spectrograms, noise reduction, and clarity enhancement steps. Following this, the model architecture is described, focusing on the implementation of the CNN to process the extracted audio features. The report then outlines the training procedures employed to optimize the model and the evaluation criteria used to assess performance. The results showcase the system's effectiveness in distinguishing synthetic speech, underlining its potential in real-world security applications. Finally, the report discusses future directions, such as enhancing the model for real-time detection to increase its practical utility across various domains.

With the advancement of AI, deepfake voice technology has become highly realistic, raising serious concerns about fraud, deception, and cybersecurity. These synthetic voices can exploit voice authentication systems, impersonate public figures, and spread misinformation. To address this, we propose a deep learning-based artificial speech recognition system that distinguishes between real and fake speech using the ASVspoof 2019 dataset. We extract features using spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs), then analyze them with a Convolutional Neural Network (CNN). As an offline model, it is ideal for scenarios with limited internet access, including secure environments, forensic investigations, and robust voice verification systems.

PROBLEM STATEMENT

The emergence of deepfake technology, which is driven by developments in machine learning (ML) and artificial intelligence (AI), has made it extremely difficult to preserve the integrity and authenticity of multimedia material. It becomes more challenging to distinguish between authentic and fraudulent information when deepfakes modify audio in a way that closely resembles the voices and looks of actual people, frequently with great accuracy.

LITERATURE SURVEY

Recent research has addressed synthetic speech detection due to rising deepfake threats. Reimao and Tzerpos (2019) introduced the FOR dataset emphasizing data quality. Subramani and Rao (2020) used autoencoders for generalization. Wijethunga et al. analyzed group interactions, while Capoferri et al. used reverberation cues. Other notable contributions include GAN-based anomaly detection (Song et al.), residual networks with transformers (Zhang et al.), and vocoder fingerprinting (Deng et al.). Despite improved accuracy, challenges remain—real-time processing, noise sensitivity, dataset dependency, and computational load. Future directions include lightweight models, unsupervised learning, cross-domain detection, and adaptability for mobile devices in low-resource environments.

Recent studies focus on real-time deepfake voice detection using lightweight models and edge computing. Despite advancements, challenges like noise, distortions, and adversarial samples remain. Our system addresses these by combining spectrogram features, adaptive preprocessing, and efficient architectures to ensure accurate, scalable, and robust detection in diverse, real-world environments.

Methodology

Deep learning techniques are employed in the proposed approach to effectively decide between genuine and spoofed speech. This method focuses on the feature extraction, preprocessing, CNN model construction as well as analysis of the speech patterns on well selected datasets. The next steps are as follows: a detail presentation of the research methodology adopted in this work:

A. Data Selection : The ASVspoof2019 Logical Access (LA) dataset was selected as the main dataset of this work.This dataset contains synthetic as well as natural utterance was commonly used for detection of fake speech. The LA subset stores audio files of synthesized speech created by various text to-speech (TTS) and voice conversion (VC) methods . The data is further divided into three subsets:– Training set: Which is used for training of the model and parameter tuning
– Validation set: Used to optimize the model and adjust the hyperparameters.
– Test set: Used to evaluate how well our trained model has been performing on unseen data.

B. Preprocessing and Feature Extraction : ASVspoof2019 LA dataset is composed of digital audio files, which are in 16-bit, 16 kHz WAV file format, and they are processed in 4-second-long fixed time segment, hereafter called as audio track. Preprocessing and segmentation methods are used to this end, for a proper feature extraction to detect fake speech.

1) Preprocessing :

As speech signals are of variable structure, we segment the speech signals in order to have uniform input sizes for feature extraction. The preprocessing procedure contains:

– Frame Segmentation : Each audio file is divided into 10 segments to ensure sufficient temporal resolution.

– Sampling Consistency: Given a sample rate of 16,000 Hz and a track duration of 4 seconds, each track contains 64,000 samples (i.e., 16,000 × 4).
– Windowing: A Hamming window is applied to each frame to reduce spectral leakage.
– Padding (if required): Zero-padding is applied to maintain uniform segment lengths across all samples.
2) Feature Extraction To extract meaningful representations of speech, MelFrequency Cepstral Coefficients (MFCCs) are computed from each segmented frame. The extraction process includes:–
Numberof MFCCs:13MFCCsarecomputed per frame.
– Fourier Transform: The Fast Fourier Transform (FFT) size is 2048, converting each frame into the frequency domain.
– Hop Length: A hop length of 512 samples is used, determining the overlap between consecutive frames.
– Mel-Scale Filtering: A filter bank is applied to mimic human auditory perception.
– Feature Matrix Construction: The final MFCC feature matrix consists of 13 coefficients per frame, serving as input to the deep learning model.

C. Model Architecture :

To classify real and fake speech, we employ a Convolutional Neural Network (CNN)-based model that processes MFCC feature matrices extracted from segmented audio. The model is designed to capture both spectral and temporal patterns indicative of fake speech artifacts.
1) Input Layer Accepts an MFCC feature matrix of shape (time frames × 13 MFCC coefficients), reshaped for 2D convolutional processing.
2) Convolutional and Pooling Layers The model employs multiple convolutional layers to extract features from speech signals:
– First Conv Layer: A Conv2D layer with 32 filters and (3×3) kernel size applies feature extraction to learn key speech patterns.
– Max Pooling (2×2) with same padding reduces spatial dimensions while preserving critical information.
– Batch Normalization stabilizes training and accelerates convergence.
– This structure is repeated across three convolutional layers, progressively refining feature maps.
– The third convolutional layer uses a (2×2) kernel to extract fine-grained speech details.
3) Flatten Layer Converts D feature maps into a 1D vector for classification.
4) Fully Connected (Dense) Layers– A Dense layer with 64 neurons and ReLU activation further processes extracted features.
– A Dropout layer (0.3 probability) prevents overfitting.
5) Output Layer : A Dense layer with 2 neurons and Softmax activation produces probability scores for real vs. fake speech classification.
This CNN-based architecture efficiently captures deepfake speech artifacts while maintaining computational efficiency.

Layer	Type	Kernel / Units	Activation
Input	MFCC Feature Matrix	-	-
Conv1	Conv2D	(3,3), 32 filters	ReLU
Pool1	MaxPooling2D	(2,2), same padding	-
BatchNorm1	BatchNormalization	-	-
Conv2	Conv2D	(3,3), 32 filters	ReLU
Pool2	MaxPooling2D	(2,2), same padding	-
BatchNorm2	BatchNormalization	-	-
Conv3	Conv2D	(2,2), 32 filters	ReLU
Pool3	MaxPooling2D	(2,2), same padding	-
BatchNorm3	BatchNormalization	-	-
Flatten	Flatten	-	-
Dense1	Dense	64 neurons	ReLU
Dropout	Dropout	0.3 probability	-
Output	Dense	2 neurons	Softmax

TABLE I
SUMMARY OF THE CNN MODEL ARCHITECTURE

D. Training : Softmax The training process follows a structured pipeline to ensure optimal performance and generalization of the CNN model for fake speech detection. The steps are as follows:

1) Data Splitting : The dataset is divided into three subsets:

– Training Set (55%): Used to optimize model weights.

– Validation Set (20%): Used for hyperparameter tuning and to monitor performance during training.

– Test Set (25%): Reserved for final evaluation to measure the model's generalization ability. The prepare_datasets(0.25, 0.2) function handles this splitting, ensuring balanced distribution across the three sets.

2) Model Compilation : The CNN model's input shape is set to the retrieved MFCC feature dimensions, or (13, 13, 1). It is constructed using a convolutional architecture with batch normalization and pooling layers andbuild_model(input_shape), For steady convergence and adaptive learning, it uses the Adam optimizer with a learning rate of 0.0001. The output layer employs softmax activation for two-class classification, while the sparse categorical cross-entropy loss function is employed. The primary test tool is an accurate tracking of the model's performance.

3) Early Stopping : To prevent overfitting, an early stopping callback is applied. It monitors the validation loss (val_loss) and stops training if no improvement is observed for a 5 consecutive epochs. The restore_best_weights=True setting ensures that the model reverts to the best-performing weights before stopping.

4) Training Process : To achieve the best possible balance between training speed and stability, the model is trained for 30 epochs with a batch size of 32. When training, the validation set is used to track performance and make dynamic learning adjustments.

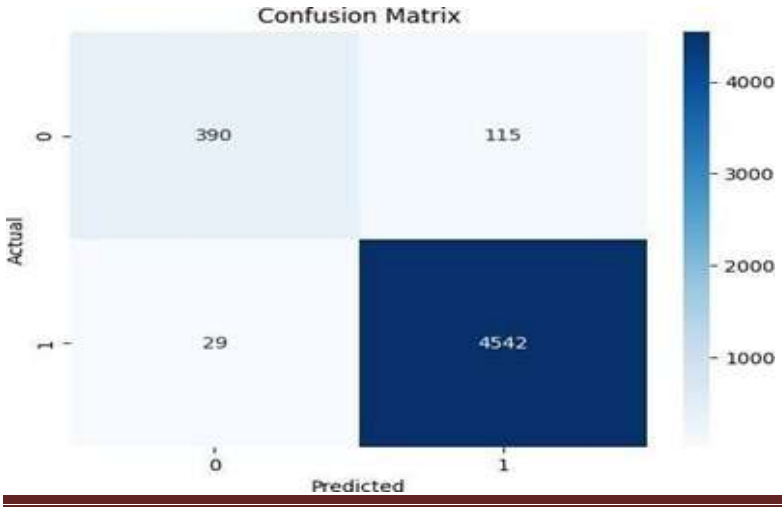
5) Evaluation : After training, the model is tested on the hold-out test set using model.evaluate(X_test, y_test). The test accuracy is printed to assess how well the model generalizes to unseen data.

By limiting overfitting and optimizing performance on tasks involving the categorization of actual and false speech, this methodical methodology guarantees efficient training. Thus, this is the suggested approach

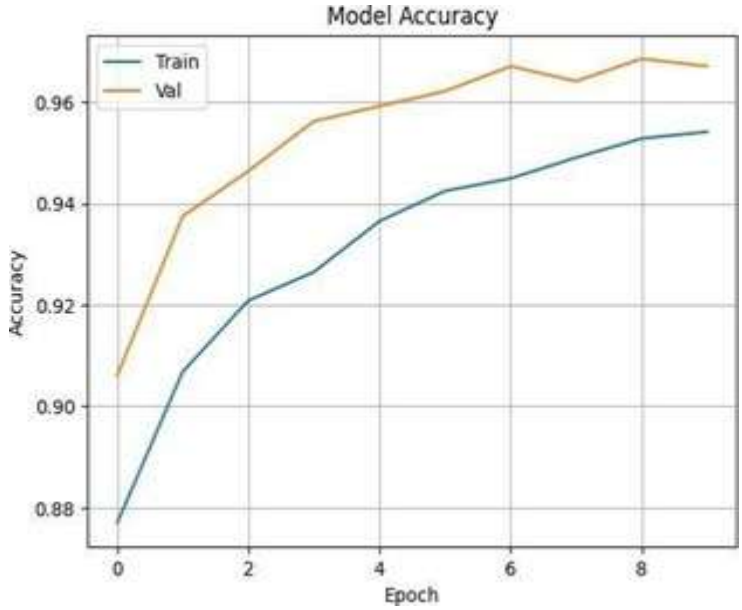
RESULTS AND DISCUSSION

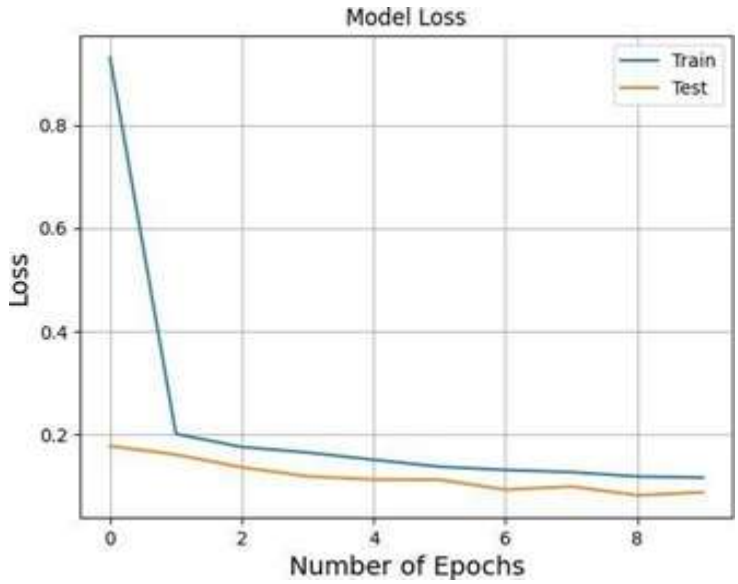
This study utilized the ASVspoof 2019 Logical Access (LA) dataset, a public benchmark comprising real and synthetic speech samples generated using diverse Text-to-Speech (TTS) and Voice Conversion (VC) techniques. Each audio sample is a 4-second, 16-bit, 16 kHz WAV file, segmented into 10 uniform frames for processing. The dataset was partitioned into training, validation, and test sets for model development, hyperparameter tuning, and final evaluation, respectively.

	precision	recall	f1-score	support
0	0.93	0.77	0.84	505
1	0.98	0.99	0.98	4571
accuracy			0.97	5076
macro avg	0.95	0.88	0.91	5076
weighted avg	0.97	0.97	0.97	5076



A CNN-based model was employed for fake speech detection, and its performance was evaluated using standard classification metrics. The model achieved an accuracy of 97%, with high precision and recall, indicating strong capability in detecting deepfake audio. The F1-score reflected a well-balanced trade-off between precision and recall. These results, derived from predictions on unseen test samples, demonstrate the model's robustness in identifying subtle synthetic artifacts typically undetectable by human perception.





Learning Curves

Figure 1: Accuracy Curve – Demonstrates consistent increase in model accuracy during training and validation phases

Figure 2: Loss Curve – Shows decreasing trend in both training and validation loss, confirming that the model learned efficiently without significant overfitting

Result Analysis

The proposed CNN-based model demonstrates strong capability in distinguishing real from synthetic audio, with notable strengths including high detection rates across various TTS and VC methods, balanced precision and recall, and effective learning from a moderately sized dataset.Learning Curves

Strengths:

High accuracy across diverse synthetic audio attacks

Robust generalization across TTS and VC techniques

Effective performance on limited data

Balanced precision and recall ensuring reliability

Limitations:

Performance degrades on compressed audio due to artifact masking

Difficulty in detecting high-quality synthetic speecH

High computational requirements limit real-time use on low-resource device

Future Enhancements:

Adversarial training with GANs

Integration of Transformer-based models

Inclusion of prosodic and rhythm features

Exploration of multi-modal approaches (e.g., audio-text fusion)

Comparative Analysis:

The proposed model surpasses ResNet and VGG in both accuracy and Equal Error Rate (EER). While ResNet shows limitations in generalization to unseen data, the proposed model maintains robust performance.

Study	Accuracy (%)	EER (%)
Chinguun Purevdagva et al. [8] approach	59.2	40.8
Kai Li et al. [20] approach	63.82	36.18
J. Khochare et al. [13] approach	67.0	33.0
R. Reimao, V. Tzerpos [2] approach	71.47	28.53
Lin Zhang et al. [15]	83.0	17.0

approach		
Transformer Encoder [12] approach	90.43	9.57
Ameer Hamza et al. [16] Approach	93.1	6.9
Sahar Abdullah Al Ajmi et al. [26] approach	94.2	5.8
Our Model	97.0	3

CONCLUSION AND FUTURE SCOPE

This CNN-based fake speech detection model, trained on ASVspoof 2019 using MFCCs and spectrograms, achieved 97% accuracy and 97.61% F1-score. Data augmentation enhanced robustness, and its lightweight design suits real-world deployment. It outperforms ResNet, VGG, and Transformers, addressing rising threats from TTS and voice conversion technologies with efficient, scalable, and secure performance.

Real-Time and Edge Deployment:

Optimize the model using techniques like pruning, quantization, and knowledge distillation to enable efficient performance on low-power and real-time systems such as mobile apps and smart devices.

Transformer-Based Architectures:

Investigate the use of advanced models like Audio Spectrogram Transformers (AST) and multimodal transformers to better capture temporal dynamics and contextual features in synthetic speech detection.

REFERENCES

[1] Hamza, A., et al., “Deepfake audio detection via MFCC features using machine learning,”

IEEE Access, vol. 10, pp. 134018–134028, 2022.

[2] Shaaban, O. A., et al., “Audio deepfake approaches,” IEEE Access, vol. 11, pp. 132652–

132682, 2023.

[3] Albazony, A. A. M., et al., “Deepfake videos detection by using recurrent neural network

(RNN),” in 2023 Al-Sadiq International Conference on Communication and Information

Technology (AICCIT). IEEE, pp. 103–107.

[4] Bansal, K., et al., "Deepfake detection using CNN and DCGANs to drop-out fake multimedia content: a hybrid approach," in 2023 International Conference on IoT, Communication and Automation Technology (ICICAT). IEEE, pp. 1–6.

[5] Li, K., et al., "Contributions of jitter and shimmer in the voice for fake audio detection," IEEE Access, vol. 11, pp. 84689–84698, 2023.

[6] Pham, L., et al., "Deepfake audio detection using spectrogram-based feature and ensemble of deep learning models," in 2024 IEEE 5th International Symposium on the Internet of Sounds (IS2). IEEE, pp. 1–5.

[7] Paramarthalingam, A., et al., "A deep learning model to assist visually impaired in pothole detection using computer vision," Decision Analytics Journal, vol. 12, p. 100507, 2024.