INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 09 ISSUE: 05 | MAY - 2025 SJIF RATING: 8.586 ISSN: 2582-3930

"Deepfake Detection in Call Recordings:A Deep Learning Solution for Voice Authentication"

Suhas Nimbalkar
Dept. of Computer Engineering,
Trinity College of Engineering & Research,Pune.
Pune,India.
suhasnimbalkar62@gmail.com

Niraj Bankar
Dept. of Computer Engineering,
Trinity College of Engineering & Research,Pune.
Pune,India.
nirajbankar3131@gmail.com

Omkar Mali
Dept. of Computer Engineering,
Trinity College of Engineering & Research,Pune.
Pune,India.
omkarmali2103@gmail.com

Sandip Bhande
Dept. of Computer Engineering,
Trinity College of Engineering & Research,Pune.
Pune,India.
saneepbhande12003@gmail.com

Dr. Geetika Narang
Dept. of Computer Engineering,
Trinity College of Engineering & Research,Pune.
Pune,India.
geetikanarang.tcoer@kjei.edu.in

***Abstract*— The emergence of deepfake technology has improved exponentially and this intensified the fears that surround the credibility of audio recordings, in instance telecommunication and security. This project proposes a full deep learning-based approach to deepfake voice recordings detection in call communications as an improvement to the voice authentication processes used. It is with this in mind that we came up with an adaptive architecture that positions convolutional neural networks (CNN) and recurrent neural network (RNN) in a way that assists in distinguishing between real and fabricated sounds. The nature of the problem allows the use of a large amount of data collected from a wide variety of real and fake audio samples which serve for proper training and testing of the system. To improve the performance of the model, some strategies have been implemented including audio preprocessing such as spectrogram and features. This research adds to the existing body of literature on voice authentication but also seeks to underscore the need for solutions that secure audio communication in times when deepfakes are on the rise. Subsequent research will be dedicated to perfecting the existing model and assessing the feasibility of its use in practice.**

Keywords*— Deepfake Detection, Voice Authentication, Synthetic Speech Analysis, Speech Forensics, Audio Deepfake Identification, MFCCs, Spectrogram Analysis, CNNs, RNNs, Transformer Models, Automatic Speaker Verification (ASV), AI Security, Fraud Prevention, Adversarial Attacks, Machine Learning for Audio Forensics.

I. INTRODUCTION

Artificial intelligence and deep learning revolutionized speech synthesis so that one can produce realistic deepfake voice. Such processes are capable of mimicking someone's voice through emulating pitch, tone, cadence, and speaking styles. Misuse is a major security concern because such voices could be exploited by cybercriminals to bypass voice verification processes, impersonate persons, dissemate disinformation, and perform fake operations.The financial industry highly depends on voice-based biometric verification, hence becoming a valuable target for deepfake-driven fraud. In countering this, this research recommends a deep learning-driven deepfake detection system developed with call recording and voice authentication systems in mind. The system applies state-of-the-art feature extraction mechanisms and deep learning frameworks to detect fine speech patterns. This powerful AI-based detection model is designed to make voice authentication systems more secure, counter deepfake cyber attacks, and assist in speech forensics and AI security.

Figure 1.1: Deepfake Detection in Call Recordings Background

Artificial intelligence and deep learning have transformed speech synthesis and voice biometrics systems, resulting in the creation of deepfake audio technology. Deepfake audio technology employs machine learning models to produce deep synthetic voices, resembling human speech. Deepfake audio technology has uses in assistive communication, dubbing films, and virtual assistants, but its misuse affects security. Fields such as banking, finance, cybersecurity, and law enforcement are exposed to deepfake attacks, which result in unauthorized transactions, fraud, and social engineering attacks. The availability of open-source tools makes it convenient for users to generate high-quality synthetic voice forgery, making the conventional speaker verification system inefficient in identifying deepfake audio attacks.

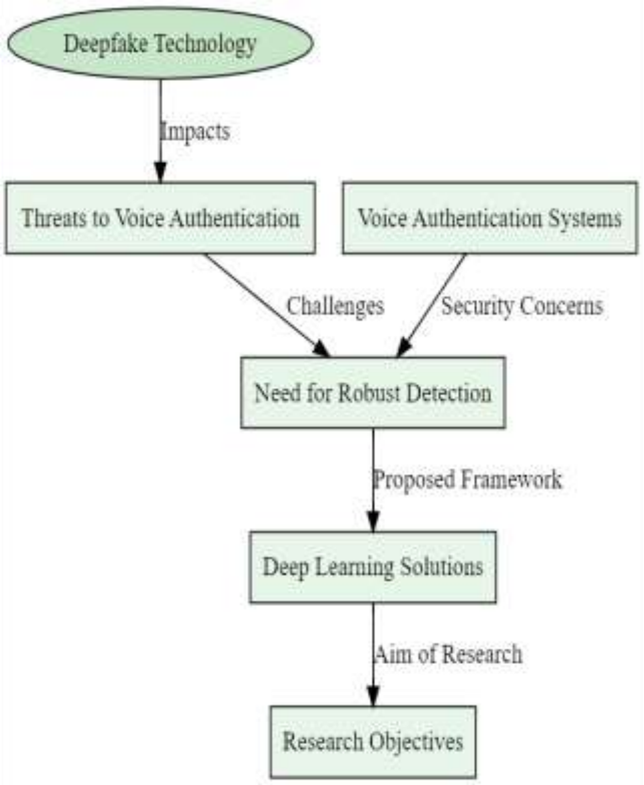


Figure 1.2: Introduction Diagram

II. LITERATURE REVIEW

Deepfake detection has emerged as a significant field of study with the fast-paced development of synthetic media generation technology. Several methods involving deep learning, signal processing, and explainable AI have been suggested to address these threats.

In [1], Abir et al. suggested a deep learning-based system for detecting deepfake images using Explainable AI (XAI) methods to enhance the model interpretability. The paper illustrates how convolutional neural networks (CNNs) can reliably separate real and fake images, and visualization tools inform the model's decision-making process.
Lim et al. [2] used this rule in the audio context by developing an explainable deep learning model to identify deepfake voices. They employed audio features such as spectrograms and integrated XAI methods for explaining the effect of modification in acoustic features on model predictions.
Yi et al. [3] presented an in-depth overview on detecting audio deepfakes and grouped current methods under signal-based, feature-based, and deep learning-based categories. They also provide key datasets, metrics, challenges, and provide useful guidance for future work on the subject.
In [4], Iqba et al. tested feature engineering techniques for deepfake audio detection using traditional machine learning models. Their strategy involves the extraction of the respective acoustic features and the use of classifiers such as SVM and Random Forest, with emphasis on the possibility of light-weighted detection systems.
Shaukatali et al. [5] have designed a spoofing voice detection scheme employing elementary audio analysis methods and machine learning classifiers. Their method aims for real-time detection with a compromise between detection accuracy and computational efficiency.
Kokate et al. [6] proposed an Exception Model based on anomaly detection for deepfake audio identification. The model is focused on acoustic anomalies in order to detect subtle manipulations that cannot be detected by ordinary classifiers, providing a novel approach in this field.
A more comprehensive review of the literature is provided by Imran et al. [7] that spans detection techniques in deepfake image, video, and audio areas. Their review presents common tools, challenges, and trends, focusing on the interdisciplinary aspect of deepfake research.
Babiker et al. [8] examined the application of deepfake voice technology in cases of scams. Their study mimics the threat in the real world through the production of synthetic voices and threat assessment, and through it offers important insights into cybersecurity and fraud prevention.
Kulangareth et al. [9] proposed a new method for deepfake voice detection through speech pause patterns. The approach is based on the temporal nature of speech and has very high accuracy in discriminating between real and fake audio.
Kotha et al. [10] wrote about machine learning-based classification of AI-generate voice. Their method focuses on good feature extraction as well as model training, proposing an efficient detection pipeline for voice modification.
Lim et al. [11] continued to further develop their explainable deep learning model for voice deepfake detection. The newer model incorporates stronger interpretability aids and better feature extraction techniques in support of even more transparent and accurate synthetic speech classification.
Malviya et al. [12] compared deep learning and machine learning models for identifying deepfakes in image, video, and audio datasets, and their work introduces the relative advantages of traditional algorithms and neural networks in a hybrid system design practical guide.
In [13], Al-Khazraji et al. discussed broader social impacts of deepfake technology, i.e., its influence on misinformation and the perception of social media users. According to them, detection technologies must be employed to fight against ethical and psychological impacts.
Munir et al. [14] created a targeted dataset of deepfake audio in Urdu to balance the absence of non-English languages from representation in detection studies. The study helps enhance multilingual AI equity and diversity of datasets to allow better detection in the local setting.
Jebamani et al. [15] proposed a traditional audio analysis-based model for fake audio detection. Their system analyzes key audio features and shows decent performance using lightweight classifiers, positioning their model as suitable for low-resource environments.

III. RESEARCH METHODOLOGY

SYSTEM DESIGN

The system design of "Deepfake Detection in Call Recordings: A Deep Learning Solution for Voice Authentication" is explained in this chapter. Data flow, system architecture, and the detailed design elements are presented to implement and deploy it.

SYSTEM ARCHITECTURE

System design is divided into several major parts to ensure efficient deepfake detection in call records. All these parts are combined to process voice input, extract certain features, and categorize the voice as fake or real.

Components

1. Audio Input Module: Stores call recordings or takes voice input. It takes multiple audio inputs like WAV, MP3, etc.
2. Preprocessing Module: The module pre-processes raw audio via noise removal, normalization, and feature extraction via techniques like MFCC (Mel-frequency cepstral coefficients), spectrogram analysis.
3. Deep Learning Model: The model makes use of a deep learning model, i.e., a set of Convolutional Neural Networks (CNN) in the feature extraction step and LSTM (Long Short-Term Memory) networks to process audio temporal patterns. The module extracts features and classifies the audio as synthetic or real.
4. Classification Module: Looks at the output from the deep learning model to either classify audio as original or deepfake. Gives a confidence level based on prediction done by the model.
5. Integration Layer: Offers smooth interconnectivity of the deepfake framework with existing voice authentication

frameworks. It handles API requests and responses in real time.

6. User Interface (UI) Module: Displays outcomes to system administrators like flagged suspicious calls and system state. The UI also facilitates system configuration and monitoring of performance.
7. Database: Exists for storing call records, detection outcome logs, and other metadata to be audited later and for further analysis.

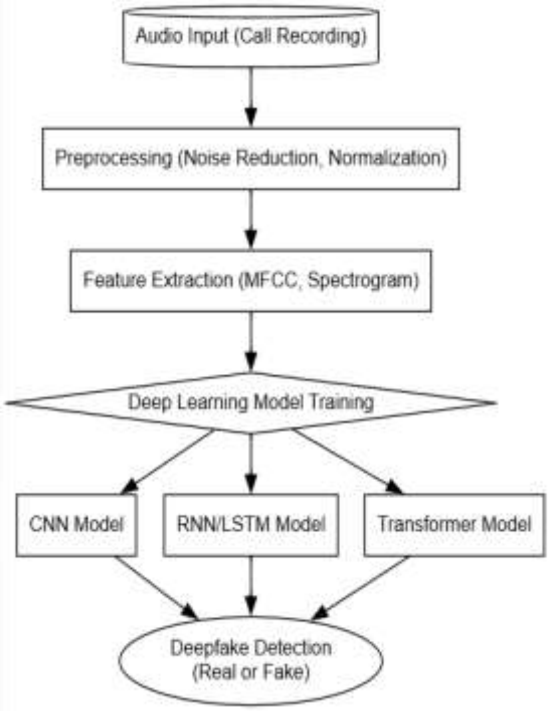

Audio Input (Call Recording)
Preprocessing (Noise Reduction, Normalization)
Feature Extraction (MFCC, Spectrogram)
Deep Learning Model Training
CNN Model
RNN/LSTM Model
Transformer Model
Deepfake Detection
(Real or Fake)

Figure 2.1: System Architecture

Algorithm Steps:

Deep Learning Model Development Overview

- Input Audio Data: Take call recordings as an input data.
- Data Preprocessing: Normalize audio to a standard volume level and segmented audio for processing purpose.
- Feature Extraction: It extracts Mel-Frequency Cepstral Coefficients (MFCCs) for voice features and builds spectrograms.
- Model Training: Splits dataset into training sets, validation sets, and testing sets.
- Model Evaluation: Tests model with the validation dataset and calculates performance metrics.
- Real-Time Detection: Implements trained model into a real-time system.
- Output Results: Displays classification outcomes in a convenient-to-use interface.
- Continuous Improvement: Gathers feedback from and updates model to adapt to evolving deepfake technologies.

IV. RESEARCH FRAMEWORK

The research follows a systematic methodology involving feature extraction, model implementation, training, evaluation, and dataset selection.

Dataset Selection: Training and validation are done using publically available datasets such as ASVspoof, FakeAVCeleb, and LJSpeech. Noise, time stretching, and pitch shifting should be added to data for simulating real-world scenarios.

Feature Extraction: Short-term power spectrum information is extracted from speech using Mel-Frequency Cepstral Coefficients (MFCCs). Spectrogram analysis transforms speech signals into visual representations for CNN-based classification. Raw Waveform Analysis directly processes audio data using deep learning algorithms.

Deep Learning Model Implementation: Convolutional Neural Networks (CNNs): Effective for classification based on spectrograms. LSTMs and Recurrent Neural Networks (RNNs) are efficient for processing voice data in a sequential manner. Transformers: Cutting-edge models for learning speech representations (Wav2Vec, Whisper, or wavLM).

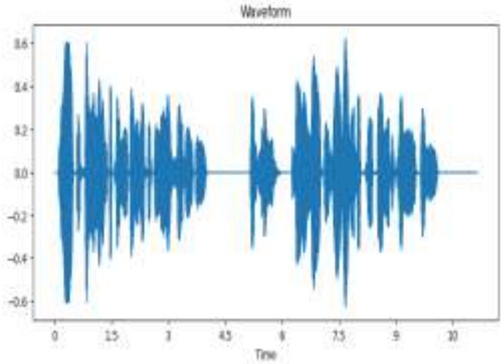
Training and Optimization: Optimize models with the Adam optimizer and cross-entropy loss function. Using grid search and random approaches to adjust hyperparameters. Evaluation Metrics: Precision, Accuracy, Recall, and F1-score, Receiver Operating Characteristic (ROC) Curve and Area Under Curve (AUC).

Evaluation Metrics: Precision, Accuracy, Recall, and F1-score, Receiver Operating Characteristic (ROC) Curve and Area Under Curve (AUC).

V. RESULT

Dataset

The dataset consists of real and fake call recordings. These call recordings are collected to train an LSTM model which helps in deepfake audio detection. Each audio file first undergoes preprocessing, including noise removal and resampling at 22,050 Hz. The feature extraction process utilizes Mel-Frequency Cepstral Coefficients (MFCC), which captures the most relevant speech features. The extracted features are then standardized and reshaped for compatibility with the LSTM model.

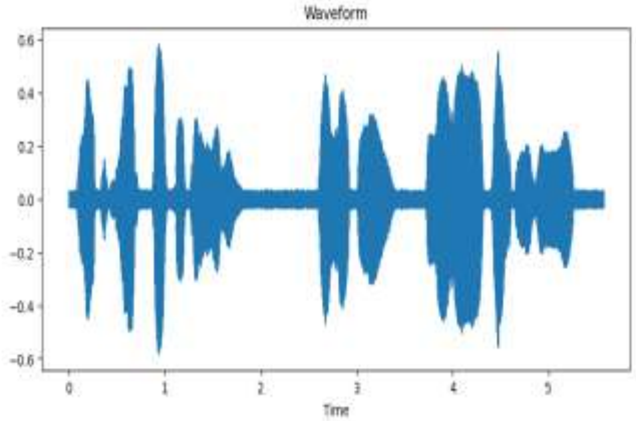


Graph 4.1: Standardized and Reshaped for Compatibility with the LSTM Model

These visualizations provide insights into the characteristics of the audio signals:

Waveform Plot

This plot shows that how the audio signal's amplitude changes over time. It facilitates comprehension of the duration and intensity of speech patterns in both synthetic and real-world calls. Deepfake audio may exhibit unnatural speech modulations due to variations in amplitude changes.

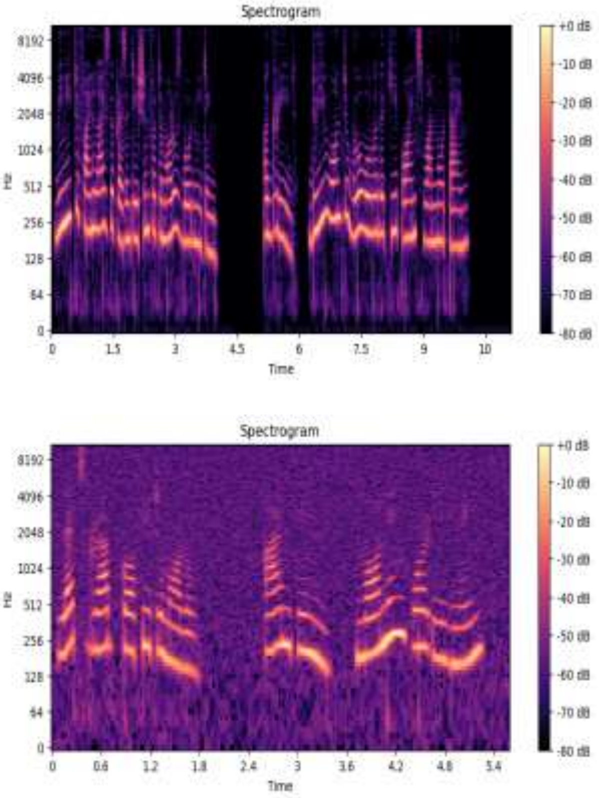


Graph 4.2: Waveform Plot

Spectrogram Plot

The spectrogram is a representation of the frequency content of the audio signal over time. It performs conversion of the signal to the frequency domain using Short-Time Fourier Transform (STFT).

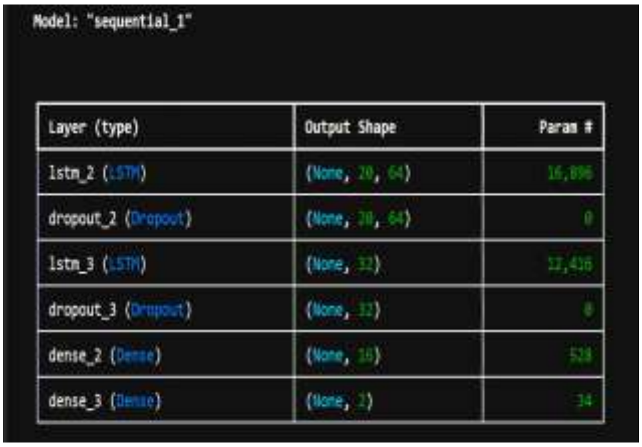
The intensity of frequencies is represented in decibels (dB) as a color-coded scale. Irregular spectral patterns, distortion, or unnatural distribution of frequencies are some of the characteristics of deepfake audio compared to natural speech.



Graph 4.3: Spectrogram Plot

Before training, plots of waveform and spectrogram for actual and synthetic audio samples exhibit differences in frequency patterns. The data is split into training (80%) and test (20%) sets for the purpose of generalization. The LSTM model is defined with several layers, two LSTM layers, dropout layers to prevent overfitting, and a softmax activation function for binary classification.

The model is trained using the Adam optimizer and categorical cross-entropy loss for 50 epochs. Performance metrics are measured by accuracy, precision, recall, and F1-score. Misclassification rates information is provided by a confusion matrix visualization. Post-training analysis includes feature importance evaluation, spectrogram comparison, and misclassified samples analysis. The results indicate the model's effectiveness, even though there can be occasional misclassifications due to background noise or speech distortions. Data augmentation and hyperparameter fine-tuning can improve it. There can be further refinement using data augmentation and hyperparameter fine-tuning.



```
Model: "sequential_1"

┌──────────────────────────┬────────────────────┬────────────┐
│ Layer (type)             │ Output Shape       │    Param # │
├──────────────────────────┼────────────────────┼────────────┤
│ lstm_2 (LSTM)            │ (None, 20, 64)     │     16,896 │
├──────────────────────────┼────────────────────┼────────────┤
│ dropout_2 (Dropout)      │ (None, 20, 64)     │          0 │
├──────────────────────────┼────────────────────┼────────────┤
│ lstm_3 (LSTM)            │ (None, 32)         │     12,416 │
├──────────────────────────┼────────────────────┼────────────┤
│ dropout_3 (Dropout)      │ (None, 32)         │          0 │
├──────────────────────────┼────────────────────┼────────────┤
│ dense_2 (Dense)          │ (None, 16)         │        528 │
├──────────────────────────┼────────────────────┼────────────┤
│ dense_3 (Dense)          │ (None, 2)          │         34 │
└──────────────────────────┴────────────────────┴────────────┘
```

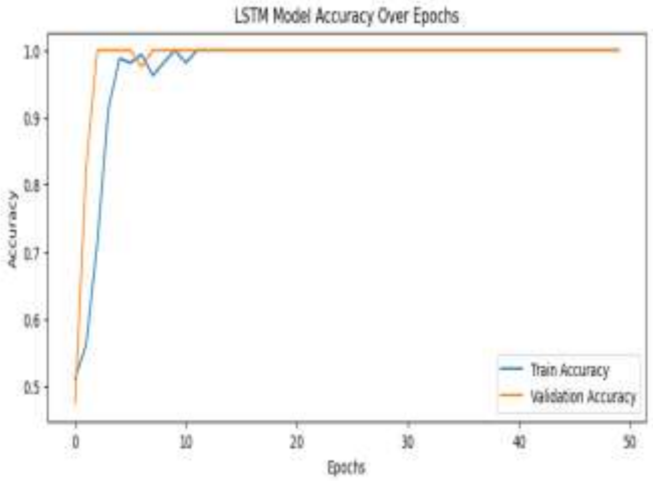
The LSTM (Long Short-Term Memory) network is structured to identify audio recordings as authentic or fabricated using derived MFCC features. The LSTM model includes two LSTM layers, in which the first 64-unit layer extracts temporal dependencies in the audio features and preserves sequence integrity, and the second 32-unit LSTM layer refines the acquired patterns.

Dropout layers of 0.2 dropout are added after every LSTM layer in order to avoid overfitting by dropping neurons randomly while training, hence allowing for good generalization. The dense layer of 16 units and activation of ReLU also aids in additional feature extraction before the final dense layer that is activated using Softmax for classification of the audio recordings into real or fake.
The model is trained with categorical cross-entropy loss, which is suitable for multi-class classification, and the Adam optimizer, which dynamically adjusts the learning rate to ensure stable convergence. Accuracy is employed as the performance measure to measure the performance of the model. The model is trained for 50 epochs with a batch size of 16, and test set is utilized for validation to approximate the generalization of the model. Training the model successfully, the model is saved as "lstm_deepfake_audio.h5" for potential reuse in testing and deployment in the future.

Result and Discussion

The LSTM model training analysis includes plotting accuracy and loss trends across several epochs. The accuracy plot indicates the trend of training and validation accuracy, giving insights into how well the model is learning the audio patterns with time. An increase in training accuracy shows effective learning, while an increasing or stable validation accuracy indicates good generalization to new data.

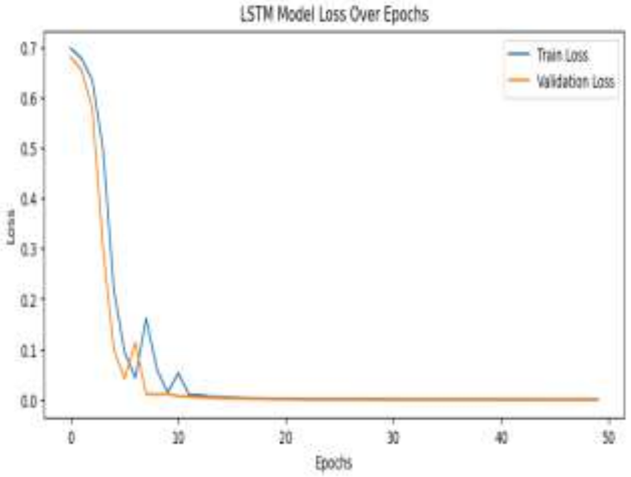
LSTM Model Accuracy Over Epochs



Graph 4.4: LSTM Model Accuracy Over Epochs

LSTM Model Loss Over Epochs

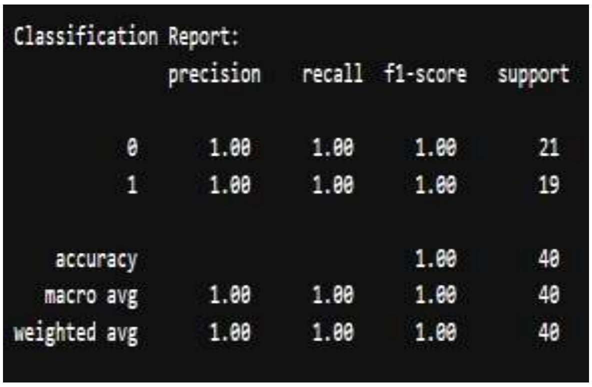
The loss plot, however, points out the way the model's categorical cross-entropy loss over epochs declines. A reducing training loss means that the model is reducing errors, while a converging validation loss means that the model is not overfitting.



If the validation loss is significantly different from the training loss, this may indicate overfitting, and adjustments like incorporating regularization, increasing dropout, or adjusting hyperparameters would be required. These plots help judge model performance, identify issues, and guide future improvements to the deepfake audio detection model.

CONFUSION MATRIX

The confusion matrix summarizes the model's performance on classifications by indicating how many correct and incorrect predictions were made for each class. Here, the matrix is comparing true and false audio predictions and can provide information about model accuracy, misclassification rates, and bias.

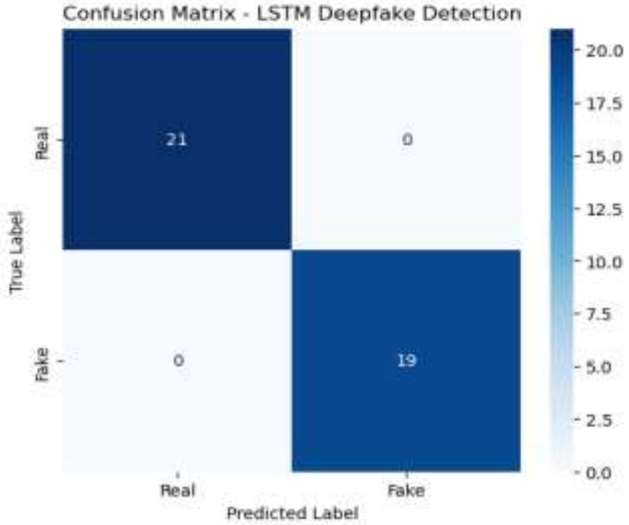


```
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        21
           1       1.00      1.00      1.00        19

    accuracy                           1.00        40
   macro avg       1.00      1.00      1.00        40
weighted avg       1.00      1.00      1.00        40
```

Good model performance is indicated by high values of correct predictions for the imposter and real samples (diagonal entries), while off-diagonal entries indicate misclassifications. Classification report also commends this study by providing precision, recall, and F1-score measures. Precision is the ratio of the predicted positive results that were indeed correct, recall is the number of actual positives correctly classified, and the F1-score is a combination of these two measures.



Graph 4.5: Confusion Matrix- LSTM Deepfake Detection

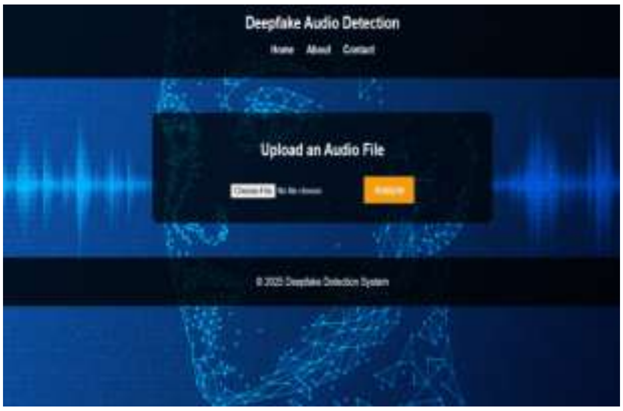
Heatmap visualization enhances interpretability by applying a color gradient to represent strengths and weaknesses of predictions. When high misclassification rates are detected, possible improvements include hyperparameter tuning, training data diversification, or using more advanced audio feature extraction techniques.

GUI (Graphical User Interface)

The Deepfake Audio Detection System has an interactive and welcoming GUI developed using Flask, which makes it possible to upload an audio file and obtain real-time classification results. Frontend is implemented using HTML, CSS, and JavaScript, and a professional-looking and modern look with a clean and responsive interface that is quite suitable on various devices. A properly selected background picture contributes to the appearance and aesthetic appeal, while a properly built navigation bar enables one to switch to any page of the application.

GUI also supports a header and footer for an organized interface, which ensures a professional and organized experience. When a user sends an audio file, the system processes the uploaded audio using the trained LSTM model, extracting the MFCC features and predicting the audio to be real or fake.

The app also includes a loading animation to inform users of processing time to ensure smoothness in the user interface. Once prediction has been done, the result is displayed in a well-presented segment with a valid label and color code to point out genuine and fake audio.



Deepfake Audio Detection

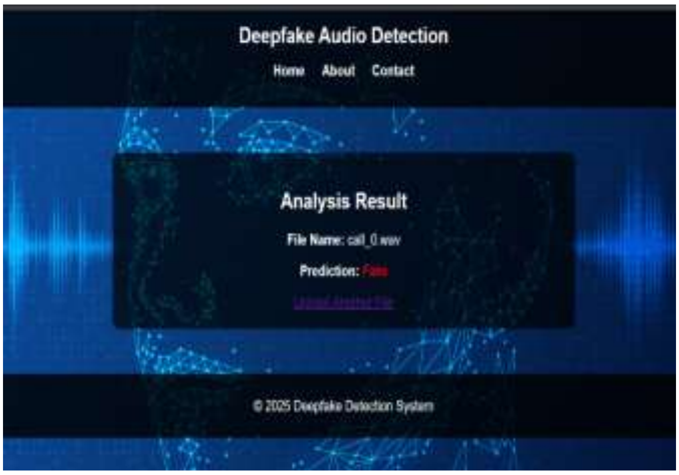
Home About Contact

Upload an Audio File

© 2025 Deepfake Detection System

Fig 4.6: Graphical User Interface

In addition, the interface has an analysis page wherein users can monitor graphical outputs such as waveform, spectrogram, accuracy, loss graph, and confusion matrix that give insights concerning the performance of the model. Generally, the design is plain, processing optimized, and easy to use.



Deepfake Audio Detection

Home About Contact

Analysis Result

File Name: call_0.wav

Prediction: Fake

© 2025 Deepfake Detection System

VI. CONCLUSION

The rapid advancement in deepfake technology challenges authentication processes based on voice severely, and hence it is imperative to identify synthetic speech for security and privacy. Here, this paper suggested a deep learning approach to identify deepfake audio from call records using neural networks to distinguish between actual and artificially generated speech.Our model, using state-of-the-art feature extraction methods and classification, attained encouraging accuracy for spoofed speech detection.

Our experiments confirm the capability of deep learning to improve voice authentication systems to provide an effective countermeasure against spoofed voice impersonation. The suggested solution, although promising, can be further enhanced with more realistic adversarial attacks, richer datasets, and multimodal verification phases. Reducing false positives and model generalization across diverse linguistic and acoustic environments are additional research areas. This work provides solid grounds for deepfake resistance in voice communication, allowing for secure and reliable authentication systems.

VII. FUTURE SCOPE

The potential of voice authentication deepfake detection is in some major breakthroughs. The real-time detection capacities can be further developed to track live calls and stop fraud in real-time. The use of adversarial training mechanisms will become even more important in order to render the defense impregnable to future sophisticated deepfake attacks. Training the detection models to be multi-language and multi-dialect compatible will render the technology world-ready and more effective and inclusive. Second, improving lightweight models for Edge AI and IoT-based deployment will make real-time security possible for mobile and embedded systems. The combination of deepfake detection and cybersecurity and digital forensics will help law enforcement and antifraud organizations prevent scams and authenticate more securely and robustly. Also, adaptive learning modules can be designed to enable AI models to continually adapt and react to novel emerging deepfake methods such that deepfakes remain barred well beyond early rollout. Such innovations will make the voice authentication platform more secure and resilient.

ACKNOWLEDGMENT

It is an earnest pleasure and a sense of satisfaction to submit this Research Paper on " *Deepfake Detection in Call Recordings: A Deep Learning Solution for Voice Authentication* " which has been made possible through the constant support, valuable guidance and meticulous direction of our guide Dr. Geetika Narang and Project coordinator Prof. Manisha Patil to whom we are deeply grateful and offer our humble thanks to Dr. Geetika Narang, H.O.D. for their constant guidance during the presentation process. The success of this Research Paper has always relied on a precise combination of hard work and unlimited co-operation and guidance, provided to us by the supervisors at our college. In addition we are thankful to our principal Dr.Abhijeet Auti whose encouragement and motivation kept us motivated and inspired us to do my best. Finally but not least, we would like to thank sincerely to colleagues, the staff and all other who directly or indirectly assisted us and made several suggestions which have undoubtedly enhanced the quality of work.

REFERENCES

[1] Suk-Young Lim et al., MDPI,2022. “Detecting Deepfake Voice Using Explainable Deep Learning Techniques”

[2] Jiangyan Yi et al., JLCF, VOL. 14, NO. 8, 2023. “Audio Deepfake Detection: A Survey”

[3] Farkhund Iqba et al., Unpaid journal. “Deepfake Audio Detection via Feature Engineering and Machine Learning”

[4] Shaikh Muskan Shaukatali et al., IJFMR, Volume 6, Issue 2,2024. “Fake Voice Detection System”

[5] Mugdha Kokate et al., IJIRSET, Volume 13, Issue 5, 2024. “Unmasking Deepfake Audio: A Study Using Exception Model”

[6] Sayed Shifa Mohd Imran et al., IRJET, Volume: 11 Issue: 03, 2024. “Deepfake Detection: A Literature Review”

[7] Ayah BAbiker et al., KTH,2024. “Deepfake Voice Implementation for Scams”

[8] Nikhil Valsan Kulangareth et al., JMIR, Vol 9, 2024. “Investigation of Deepfake Voice Detection Using Speech Pause Patterns: Algorithm Development and Validation”

[9] Mohan Krishna Kotha et al., IJCRT, Volume 12, Issue 3 ,2024. “Classification Of AI Generated Speech for Identifying Deepfake Voice Conversions”

[10] Joel Frank et al., Unpaid Journal,2019. “WaveFake: A Data Set to Facilitate Audio Deepfake Detection”

[11] Suk-Young Lim et al., MDPI, Volume 12, Issue 8, 2023. “Detecting Deepfake Voice Using Explainable Deep Learning Techniques”

[12] Riya Malviya et al., Unpaid Journal. “Deepfake Detection Using Machine Learning & Deep Learning”

[13] Samer Hussain Al-Khazraji et al., EPSTEM, Volume 23,2023. “Impact of Deepfake Technology on Social Media: Detection, Misinformation and Societal Implications”

[14] Sheza Munir et al., 2024. “Deepfake Defense: Constructing and Evaluating a Specialized Urdu Deepfake Audio Dataset”

[15] S. Anitha Jebamani et al., Ilkogretim, Vol 19 /Issue 4,2020. “Detection Of Fake Audio”

[16] Mohan Krishna Kotha et al., IJCRT, Volume 12, Issue 3,2024. “Classification Of AI Generated Speech for Identifying Deepfake Voice Conversions”

[17] Yan Ju et al., CVF. “Improving Fairness in Deepfake Detection”

[18] Kalaivani N et al., IARJSET, Vol. 11, Issue 4,2024. “Fake video detection using deep learning”

[19] Zeina Ayman et al., JCC, Vol.2, No.2,2023. “Deepfake: A Deep Learning Approach for Deep Fake Detection and Generation”