International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Deepfake Audio Detection Using MFCC Features

Priya N V[1], Pavan H[2], Prajwal S[3], Varun R[4], Vinay A[5]

[1]Assistant Professor, Banglore Institute of Technology
[2,3,4,5]Student, Banglore Institute of Techonology

Abstract

The proliferation of deepfake audio technologies poses significant threats to privacy and security, as AI-generated voices can convincingly mimic real human speech, leading to potential misuse in identity theft, fraud, and misinformation campaigns. This project aims to develop a robust system for detecting deepfake audio by leveraging advanced machine learning algorithms and signal processing techniques. The system will extract key features from audio recordings, such as Mel-Frequency Cepstral Coefficients (MFCCs), and utilize a Random Forest classifier to differentiate between genuine and manipulated audio. By analyzing subtle inconsistencies in the audio signals, the system can accurately identify deepfake content, thereby enhancing the integrity of digital communications. The project encompasses data collection of both real and AI-generated audio samples, preprocessing to extract meaningful features, and model training to achieve reliable detection results. Additionally, a user-friendly interface will be developed to allow users to upload audio files, select the level of analysis, and receive detailed detection outcomes. This initiative not only addresses the pressing need for effective deepfake audio detection tools but also contributes to safeguarding against the misuse of synthetic audio technologies in various sectors, including secure communications, media authentication, and forensic investigations.

CHAPTER 1
INTRODUCTION

The rise of deepfake technologies has made detecting manipulated audio recordings increasingly challenging, posing significant risks to privacy and security. This project aims to develop a robust system for detecting deepfake audio using advanced machine learning algorithms and signal processing techniques. By identifying subtle inconsistencies in audio recordings, our system can accurately differentiate between genuine and fake audio, helping to combat the misuse of synthetic audio technologies and ensure the integrity of digital communications. The deepfake audio detection system is a sophisticated tool designed to significantly enhance the security and authenticity of voice communications.

By leveraging advanced machine learning and audio analysis techniques, the system detects and identifies deepfake audio, ensuring that only genuine recordings are trusted. For scenarios where basic verification is required, the system quickly identifies obvious signs of manipulation, providing a straightforward assessment. For more critical or high-security applications, the system conducts a more comprehensive analysis, scrutinizing the audio for subtle anomalies that may indicate deepfake manipulation. This customization allows users to choose the level of scrutiny that best suits their needs, whether for routine checks or in-depth verification .

Problem Definition

The misuse of AI-generated deepfake audio poses significant risks in areas like identity fraud, misinform-

ation, and privacy breaches. These synthetic voices can mimic real human speech convincingly, making it difficult to differentiate between real and fake audio. The problem lies in the lack of accessible tools to reliably detect such deepfakes. This project addresses this issue by developing a machine learning-based system that analyzes audio files to classify them as real or AI-generated, offering a practical solution for fraud prevention and media verification.

CHAPTER 2
LITERARTURE SURVEY

The purpose of this section is to provide a literature, research, and scholarly works done on the topic online preparation and challenges.

2.1 Detecting Deepfakes in Audio Using Spectrogram Analysis

Authors: Korshunov, Pavel; Marcel, Sébastien

Publication: IEEE, VOLUME 11, 2020

This paper addresses the problem of detecting deepfake audio, where synthetic audio recordings are generated using advanced neural networks to mimic real human speech. The authors propose a method based on analyzing spectrograms of audio signals to detect such manipulations. Spectrogram analysis is employed because it represents the audio signal in a time-frequency domain, which is effective in capturing artifacts introduced by deepfake audio generation processes. The paper demonstrates the effectiveness of this approach through extensive experiments and provides insights into the characteristics of deepfake audio that can be exploited for detection.

2.2 Enhancing Speaker Verification for Deepfake Voice Detection

Authors: Wang, Li

Publication: IEEE 2021

This paper presents an approach to improving speaker verification systems for the detection of deepfake voices. The study focuses on enhancing the robustness of speaker verification against synthetic voice manipulations, which are increasingly used to deceive automated systems and humans alike. The authors propose novel techniques to strengthen the verification process by incorporating advanced features and models that can better distinguish between genuine and deepfake voices. The paper provides a comprehensive evaluation of the proposed methods, demonstrating their effectiveness in various scenarios and datasets. This research aims to advance the field of voice biometrics by offering improved tools for identifying deepfake audio and ensuring the integrity of voice-based systems.

3 Ethical Challenges in Deepfake Voice Detection Technology

Authors: Khan, Sarah

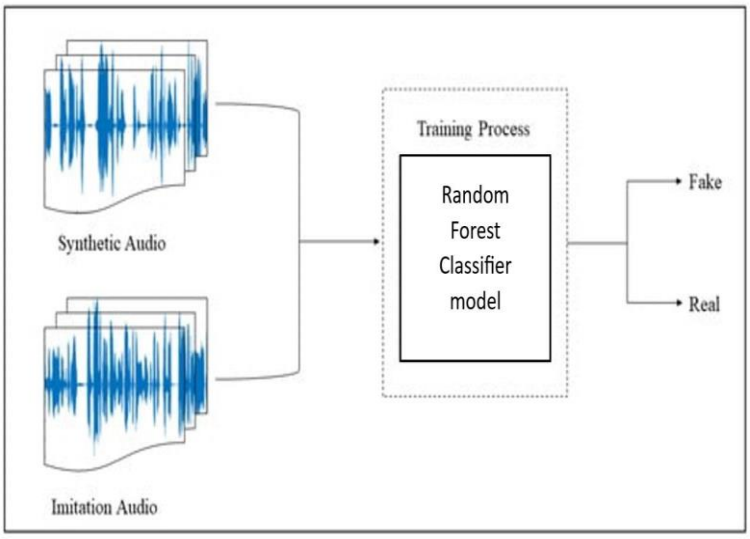
Publication: IEEE Transaction on Information Forensics and Security, 2023

As deepfake technology advances, the ethical implications of its detection and use become increasingly significant. This paper explores the ethical challenges associated with deepfake voice detection technology. It addresses issues such as privacy concerns, the potential for misuse, and the impact of false positives and negatives on individuals and organizations. The authors discuss the balance between security and privacy, the responsibility of developers and users, and the potential societal consequences of implementing detection systems. By examining real-world scenarios and case studies, the paper provides a framework for addressing these ethical concerns while advancing the field of deepfake detection. This research aims to inform policy-making and guide the development of ethical standards in the deployment of deepfake detection.

CHAPTER 3
SYSTEM DESIGN

3.1 Architectural Design

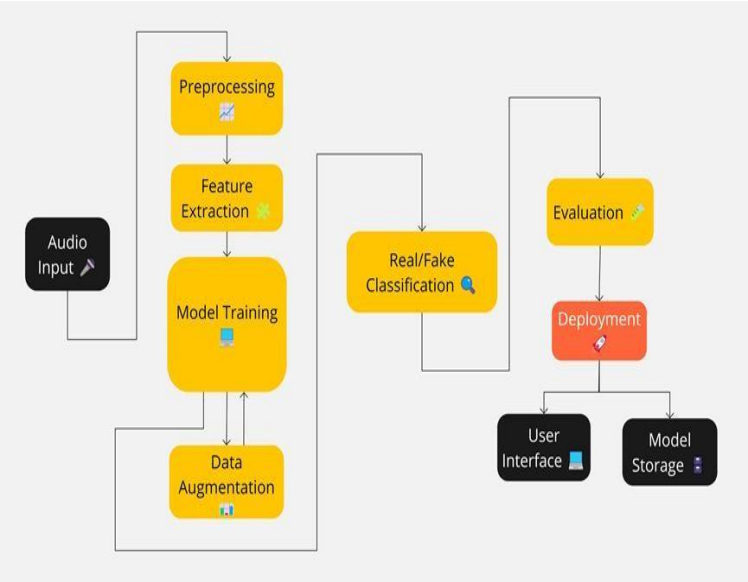
Architectural design refers to the high-level structure of a system, providing a blueprint that outlines the organization and interactions between various components. This architectural design provides a high-level overview of the deepfake audio detection system. It highlights the key stages, from input audio collection to training the detection model and finally classifying the audio. The simplified depiction underscores the importance of the training process in developing a robust model capable of accurately distinguishing between real and fake audio. By focusing on the training and classification components, this design emphasizes the core functionality of the system in detecting deepfake audio. This architectural design of your deepfake audio detection project showcases the essential components and their interactions in the process of distinguishing real audio from fake audio.



3.1 Architectural design

3.2 Detailed Design

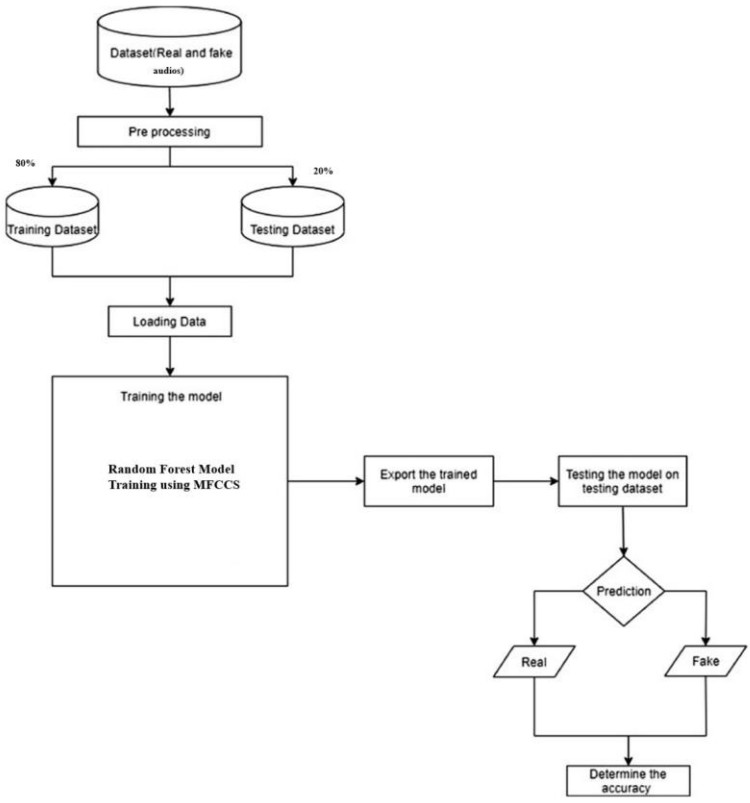
Detailed design delves into the specifics of how each component of the system will be implemented, focusing on the inner workings of both the frontend and backend. It defines the main components of the system, their relationships, and how they work together to achieve the desired functionality. In the context of the Deepfake Audio Detection project, this detailed design represents a comprehensive approach to detecting deepfake audio, integrating various stages of audio processing and feature extraction with sophisticated machine learning models. The goal of the system is to accurately distinguish between real and fake audio, which is crucial in contexts where the authenticity of audio data is of paramount importance, such as in forensic investigations, media verifications, and security systems. The system is designed to handle two primary types of audio input: synthetic audio and imitation audio. Synthetic audio refers to speech generated by text-to-speech systems or other synthetic speech generation methods. Imitation audio, on the other hand, involves audio that closely mimics genuine human speech, often created using advanced deepfake techniques. By processing these inputs through a series of extraction and classification steps, the system aims to identify subtle cues and features that differentiate authentic audio from manipulated or generated content.



3.2 Detailed Design

3.3 Data Flow Diagram

A Data Flow Diagram (DFD) is a graphical representation of the data flow through a system, illustrating how data is processed and transferred between different components. This data flow diagram highlights a multi-stage process for deepfake audio detection, utilizing both a CNN and a local model for thorough analysis. The diagram emphasizes the importance of feature extraction, model application, and decision points to ensure accurate classification of audio as real or fake. This layered approach enhances the reliability of the detection system, making it robust against various types of deepfake audio manipulations.

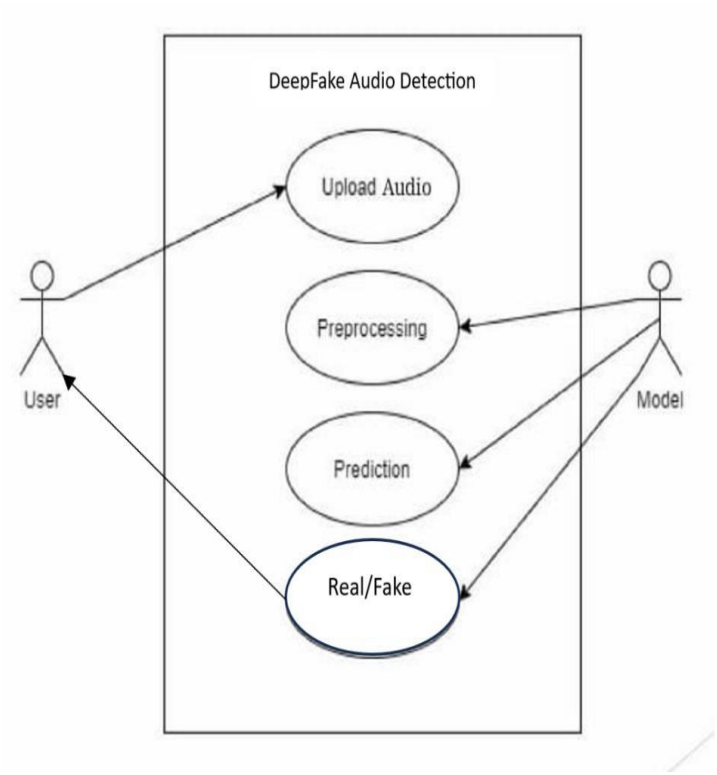


3.3 Data flow

- **Input Dataset (Real and Fake Audios):**
  - The process begins with a dataset that contains both real human audio recordings and fake (AI-generated) audio. This dataset is the foundation of the system and serves as the input for the entire pipeline.
- **Preprocessing:**
  - **Objective:** Transform raw audio into a format suitable for model training.
  - **Steps Involved:**
    - Noise removal to clean the audio signals.
    - Normalization of audio to ensure consistent levels.
    - Feature extraction (MFCCs) to convert audio signals into numerical representations.
  - Preprocessing ensures the data is structured and optimized for model training.
- **Dataset Splitting:**
  - The pre-processed dataset is divided into:
    - **Training Set (80%):** Used to train the machine learning model.
    - **Testing Set (20%):** Held back to evaluate the model's performance after training.
- **Loading Data:**
  - Both the training and testing datasets are loaded into the machine learning pipeline to begin the modeling process.
- **Model Training:**
  - **Algorithm Used:** A Random Forest Classifier, a robust ensemble learning technique.
  - **Training Details:**
    - The model learns patterns and features (like MFCCs) that differentiate real audio from fake audio.
    - The training process involves fitting the model on the training dataset and adjusting internal parameters to minimize prediction errors.
- **Export Trained Model:**
- After training, the optimized model is saved or exported for deployment.
- This model is ready to classify new audio samples into Real or Fake categories.
- **Testing the Model on the Testing Dataset:**
  - The trained model is evaluated using the testing dataset.
  - Predictions are made for each sample, and the results are compared to the actual labels to measure accuracy.
- **Prediction and Classification:**
  - For each audio sample:
    - **Prediction:** The model analyses the input features and outputs a classification.
    - **Real or Fake:** Based on the prediction, the sample is categorized as real human voice or AI-generated/fake audio.
- **Determine Accuracy:**
  - The system calculates key performance metrics:
    - **Accuracy:** The proportion of correctly classified samples.
    - Other metrics, like precision, recall, and F1-score, are used to understand the model's performance in greater depth.

3.4 Use Case Diagram

A use case diagram provides a visual representation of the interactions between users (actors) and the system, highlighting the different functionalities that users can perform. This use case diagram illustrates the workflow of a deepfake audio detection system. The process begins when a user uploads an audio file to the system, which then enters a preprocessing phase where the audio data is cleaned and prepared for analysis. The preprocessed data is fed into a predictive model designed to distinguish between real and fake audio. After prediction, the system outputs a confidence percentage, indicating the likelihood of the audio being genuine or manipulated. This result helps the user make informed decisions regarding the authenticity of the audio content. The model is central to the process, responsible for both preprocessing and generating predictions.



3.4 Use case diagram

- **Actors:**
- User: Interacts with the system to upload audio files and receive results.
- Model: The AI/ML model that processes the audio file to detect deep fakes.
- **Use Cases:**
- Upload audio: The user uploads an audio file to the system.
- Pre processing: The system preprocesses the uploaded audio (e.g., noise reduction,
  - feature extraction).
- Prediction: The model analyzes the audio file and predicts whether it is a deep fake or real.
- Fake/Real : The system returns the prediction.
- **Interactions:**
- User -> Upload audio: The user uploads an audio file.
- System -> Preprocessing: The system preprocesses the audio file.
- System -> Prediction: The preprocessed audio is fed into the model for analysis.

- System -> Fake/Real : The model returns the prediction (fake or real) to the user.
  - Fake flash real -> User: Prediction is returned to the user
- **Description:**
- Upload audio: The user uploads an audio file using a web interface or application. The
  - system receives the audio file for further processing.
- Preprocessing: The system processes the audio to clean it and extract relevant features.
- This step may involve noise reduction, normalization, and feature extraction(e.g., Melfrequency cepstral coefficients (MFCCs)).
- Prediction: The preprocessed audio is passed to the deep fake detection model. The
  - model analyzes the features and predicts if the audio is real or fake.
- Fake/Real :The system returns the prediction result to the user. The result includes whether the audio is real or fake.

CHAPTER 4
CONFUSION MATRIX

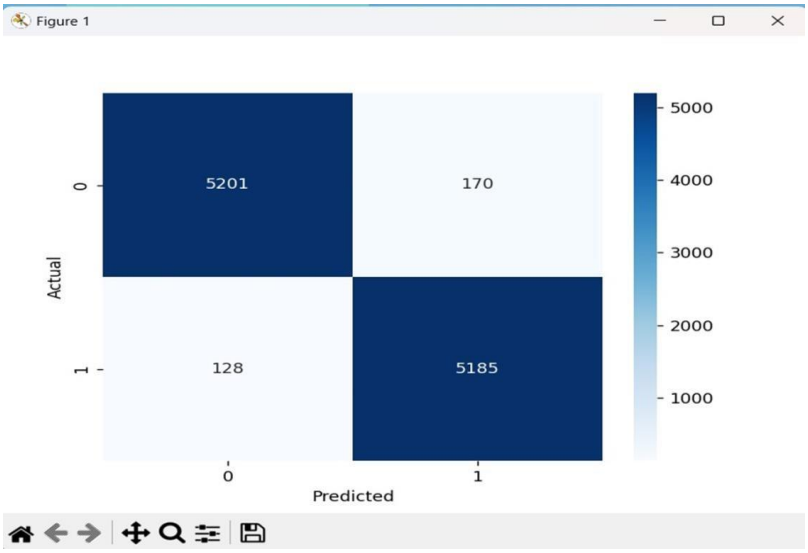


Fig 4.1 - Confusion Matrix

A confusion matrix is a performance measurement tool for machine learning classification problems. It is a specific table layout that allows you to visualize the performance of an algorithm. It summarizes the predicted and actual classifications and provides insight into the model's accuracy and error types.
The provided image shows a confusion matrix, which is a visualization of model's performance. Rows represent actual labels (ground truth), and columns represent predicted labels.

Metrics Derived from the Confusion Matrix

From the counts in the confusion matrix, you can derive several important metrics:

- Accuracy: The overall correctness of the model. Accuracy=TP+TN/TP+TN+FP+FN
- Precision: The proportion of positive identifications that were actually correct. Precision=TP/TP+FP
- Recall (Sensitivity): The proportion of actual positives that were correctly identified. Recall=TP/TP+FN

F1 Score: The harmonic mean of precision and recall, useful for imbalanced classes. F1=2×Precision×Recall/Precision+Recall

The matrix has four key values:

1. **True Positives (TP - 5185):** The model correctly predicted class 1.
2. **True Negatives (TN - 5201):** The model correctly predicted class 0.
3. **False Positives (FP - 170):** The model incorrectly predicted class 1 when it was class 0
4. **False Negatives (FN - 128):** The model incorrectly predicted class 0 when it was class 1.

This confusion matrix indicates that the classifier has high accuracy, as the majority of predictions are correct, with relatively low numbers of FP and FN. It reflects the effectiveness of the model in distinguishing between real and fake audio.

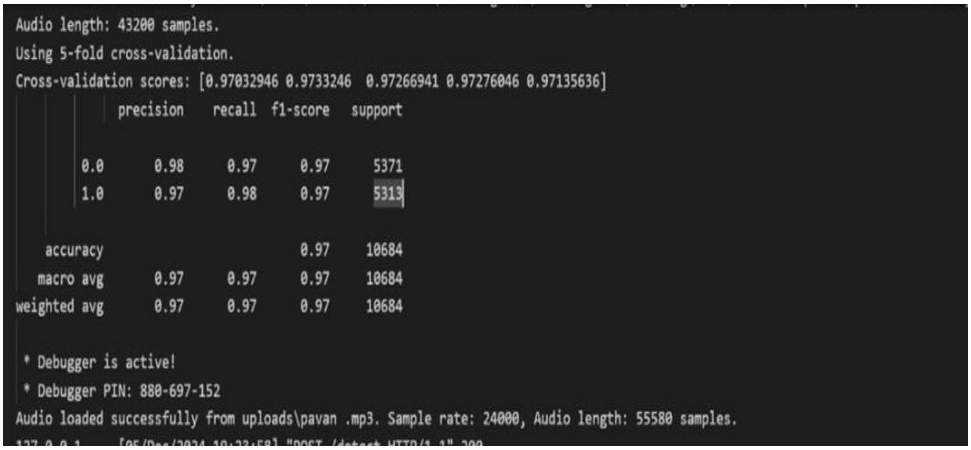


Fig 4.1-Classification Report

The Deepfake Audio Detection output highlights the model's strong performance with a 97% accuracy and balanced evaluation metrics across real (Class 0) and fake (Class 1) audio. Using 5-fold cross-validation on 43,200 audio samples, the dataset ensures robust evaluation. Precision and recall for both classes are consistently high (0.97–0.98), indicating effective detection. A balanced dataset with 5,371 real and 5,313 fake samples further supports reliability. Additionally, the Flask server successfully processed an audio file (pavan.mp3), confirming the application's proper functionality.

CHAPTER 5
RESULTS

The live recording output of the Audio Detection System showcases its ability to detect whether an audio input is real or fake. Users can record audio in real time through the interface. Once the audio is processed, the system analyses it using MFCC feature extraction and classification through the Random Forest model. The result is displayed in a user-friendly popup, providing immediate and clear feedback on the audio's authenticity.

Fig 5.1- Live Recording Output

Fig 5.2- Uploaded Audio Output

The output displayed here is the result of analysing an uploaded audio file through the deepfake audio detection system. The system processes the audio and indicates that the analysed audio file was classified as synthetic or manipulated, emphasizing its potential as fake. The text output provides a clear and concise indication of the detection system's verdict.

CHAPTER 6
CONCLUSION

The Deep fake Audio Detection Project serves as a significant step toward combating the growing threat of AI-generated audio manipulation. By employing advanced feature extraction techniques, such as MFCCs (Mel Frequency Cepstral Coefficients), and a robust Random Forest Classifier, the project effectively identifies whether an audio sample is real or fake. The system's ability to process both uploaded audio files and real-time recordings provides versatility, addressing practical needs in areas like fraud prevention, media authentication, and forensic investigations. Furthermore, the user-friendly Flask-bas1ed web application ensures accessibility for users, enabling seamless interaction and ease of use. Through its innovative approach, the project provides a scalable and efficient framework for identifying deep fake audio, catering to both technical and non-technical audiences.

This project also highlights the potential for addressing critical challenges in a digital-first era where the

misuse of AI technologies, such as deep fakes, can severely impact trust, security, and privacy. By delivering high accuracy and reliability, the project lays the groundwork for more advanced solutions, contributing to ongoing efforts in AI and cyber security research.

Collecting audio datasets in the wild: Most of the audio deepfake detection datasets are not collected in the wild, which do not quite match with the real utterances recorded or generated in realistic conditions. The real conditions of the utterances may be even worse and vary more greatly than the simulated conditions. In order to assess audio deepfake detection methods in practical applications, the utterances with a variety of channels or conditions should be collected through realistic environment conditions, such as social media platforms, Internet or telephone channels.

CHAPTER 7
REFERENCES

1. Yuezun Li, Xin Yang, Pu Sun, Honggang Qi and Siwei Lyu, Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics.
2. G. Oberoi, *Exploring DeepFakes*, [online] Available: https://goberoi.com/exploring-deepfakes-20c9947c22d9.
3. J. Hui, *How Deep Learning Fakes Videos (Deepfake) and How to Detect it*, [online] Available: https://medium.com/how-deep-learning-fakes-videos-deepfakes-and-how-to-detect-itc0b50fbf7cb9.
4. I. Goodfellow, J. P. Abadie, M. Mirza, B. Xu, D. W. Farley, S. Ozair, et al., "Generative adversarial nets," *Proc. 27th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, vol. 2, pp. 2672-2680, 2014.
5. H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, et al., "Deep video portraits", *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1-14, Aug. 2018.
6. A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Nießner, "FaceForensics: A large-scale video dataset for forgery detection in human faces", *arXiv*, 2018.
7. Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, “Spoofing and countermeasures for speaker verification: A survey,” Speech Communication, vol. 66, pp. 130–153, 2015.
8. M. R. Kamble, H. B. Sailor, H. A. Patil, and H. Li, “Advances in anti-spoofing: from the perspective of asvspoof challenges,” Published online by Cambridge University Press, pp. 1–18, 2020.
9. J. Yang and R. K. Das, “Long-term high frequency features for synthetic speech detection,” Digital Signal Processing, vol. 97, p. 102622, 2019.
10. Xue, C. Fan, Z. Lv, J. Tao, J. Yi, C. Zheng, Z. Wen, M. Yuan, and S. Shao, “Audio deepfake detection based on a combination of F0 information and real plus imaginary spectrogram features,” in DDAM@MM 2022: Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia, Lisboa, Portugal, 14 October 2022. ACM, 2022, pp. 19–26.
11. Y. Hong, Z. H. Tan, Z. Ma, and J. Guo, “Dnn filter bank cepstral coefficients for spoofing detection,” IEEE Access, vol. 5, no. 99, pp. 4779–4787, 2017.
12. N. Sainath, B. Kingsbury, A. rahman Mohamed, and B. Ramabhadran, “Learning filter banks within a deep neural network framework,” 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 297–302, 2013.
13. F. Alegre, R. Vipperla, A. Amehraye, and N. W. D. Evans, “A new speaker verification spoofing countermeasure based on local binary patterns,” in Interspeech, 2013.
14. M.Delgado, and N. Evans, “A new feature for automatic speaker verification antispoofing: Constant q cepstral coefficients,” in Processings of Odyssey 2016, 2016

15. B. Chettri, D. Stoller, V. Morfi, M. Ram´ırez, and B. L. Sturm, “Ensemble models for spoofing detection in automatic speaker verification,” in Interspeech, 2019.
16. Y . Qian, N. Chen, and K. Yu, “Deep features for automatic spoofing detection,” Speech Commun., vol. 85, pp. 43–52, 2016