National College of Ireland

ENHANCING VOICE AUTHENTICATION SYSTEMS WITH DEEPFAKE AUDIO DETECTION

MSc Research Project
MSc Cyber Security

Atharva Lawate
Student ID: X22213325

School of Computing
National College of Ireland

Supervisor: Eugene Mclaughlin

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: ……Atharva Lawate……

Student ID: ……x22213325……..……

Programme: …MSC Cyber Security…… **Year:** …2023-24……..

Module: …MSc Research Practicum…….……

Supervisor: …Eugene Mclaughlin…….……

Submission Due Date: …September 16, 2024…….……

Project Title: …ENHANCING VOICE AUTHENTICATION SYSTEMS WITH DEEPFAKE AUDIO DETECTION…

Word Count: …6690…… **Page Count**…22……..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: …Atharva Lawate……

Date: …September 16, 2024……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

ENHANCING VOICE AUTHENTICATION SYSTEMS WITH DEEPFAKE AUDIO DETECTION

Atharva Lawate

X22213325

ABSTRACT

This Report initiates a broad study on the advance of voice authentication systems with state-of-the-art mechanisms for the detection of deepfake audio. Motivated by the tremendous threat from deepfake technology in the approach of voiceprint-based security systems, the research aims to estimate and advance the efficiency of the prevailing detection procedures. Using a dataset from Kaggle and carrying out the analysis of the LSTM network on the python platform, the research also makes use of feature extraction robustness and careful evaluation metrics that are followed rigorously. Some of the key findings include the high confidence and robustness of the LSTM model to identify whether the audio is real or synthetic. These make reasonable contributions to speech and audio security, intended to be quite promising for effective work and commercialization.

Keywords - LSTM,CNN,Deep learning,Voice authentication,machine learning

1. Introduction

1.1 Background and Motivation

In the context of state-of-the-art technologies, deepfake still emerged as a powerful tool that can produce very convincing and reasonable synthetic audio and is a threat to all security systems. So lately with the rise of deep fake audio, it has become a big concern to Voice Authentication systems since it functions on the differentiation of people by their voice and the genuineness of the voice aspects. As voice-based systems are slowly spreading from banking to the security of personal devices, deepfake audio needs efficient detection solutions. Deepfake audio is one of the developments of deepfake technology in which new trends of machine learning and artificial intelligence are utilized for creating voices that are real-like. This capability not only jeopardizes the effectiveness of voice authentication systems but also provokes potential adversaries to perform actions such as identity theft or access to forbidden resources. This is why the development of proper methods to combat these threats has only been getting more attention in recent years. For instance, Chen et al. came up with a new framework that incorporates large margin cosine loss, as well as few frequency masking to increase the detection rate of the said object. Almutair and Elgibreen (2022) conducted another massive survey indicating the need for designing general and large-scale detection approaches.

1.2 Research Questions and Objectives

This report aims to address the following research questions:

1. What has to be done with the already existing technologies of voice authentication to receive and avoid deep fake audio?
2. Which of the approaches is efficient and has a higher possibility of the further development of deep fake audio detection?

To answer these questions, the research objectives include:

- To what extent, the methods existing in deepfake audio detection could be evaluated?
- To compare the pros and cons of various models of machine learning addressing the issue of detecting deepfake audio.
- To develop a new approach or adjusting the existing ones to enhance the detection rate and relevance.

1.3 Structure of the Report

The report is structured as follows:

● Related Work: This chapter brings forth the literature review on the current works done on deepfake audio detection along with the findings of the works.

● Methodology: In this section, the details of the study methodology which consists of data collection, data preparation, data feature extraction, and pattern recognition are discussed.

● Implementation: Also describes the results obtained from the experiments performed and also a comparison of all the different detection methods and the respective models.

● Discussion: Specializes in the assessment of the outcomes, elaboration of the results' significance, as well as in the suggestion of further research questions.

● Conclusion: Reiterates the findings of the study, what the research adds to the knowledge base and what needs to be done to further the research.

This report outlines the challenges and possible directions on how to improve voice authentication systems with deep fake.

2. Related Work

2.1 Introduction

Deepfake technologies have been developed in the last few years so far that it has become a real challenge for voice-based systems. Chen et al. (2020) further proposed a new framework with the use of large-margin cosine loss and frequency masking to improve the detection of audio deepfakes, which resulted in a remarkable reduction of the Equal Error Rate(EER) on the ASVspoof 2019 dataset. Almutair and Elgibreen (2022) contributed a comprehensive survey on audio deepfake detection methods, indicating accuracy at the cost of scalability and the need for more generalizable techniques. Yan et al. (2022) designed a ResNet-34-based Audio Deepfake Detection(ADD) 2022 challenge detection system that puts special attention to new techniques like neural stitching. Mcuba et al. (2023) benchmarked several state-of-the-art CNNs for deepfake detection and showed that often models specifically designed for this task perform better than general-purpose approaches. Hamza et al. (2022) investigated methods for feature extraction techniques and various machine learning models, showing the efficiency of Mel-Frequency Cepstral Coefficients and the strength of each model in different scenarios.

2.2 Empirical Study

Chen et al., (2020) discuss in their paper and consider "Generalization Of Audio Deepfake Detection" a major step towards the detection of audio deepfakes, which remains one of the most important challenges facing voice-based authentication systems today. The authors have come up with a new framework that integrates the large margin cosine loss(LMCL) function with the online frequency masking technique to improve the robustness in the detection of audio deepfakes. This research serves to answer one important question: the poor generalization of most existing methods of detection on new spoofing techniques with further development of speech synthesis and voice conversion technologies. Being significantly different from traditional methods that relied much on handcrafted features and static data, the authors' solution is integrated with LMCL and frequency masking. While LMCL can boost the neural network's discriminative power by increasing inter-class variance and decreasing intra-class variance, frequency masking is the augmentation operation that mostly increase the size of the training set to make the model more robust under different conditions. The results of this study are impressive, and the proposed system is expected to hugely reduce the EER from 4.04% to 1.26% on the ASVspoof 2019 logical access dataset. This has been achieved through several well-thought-out enhancements that involve data augmentation with background noises and simulation of the telephony environment to make the system more robust under different conditions. However, this emphasis on incremental improvements, potentially shadows the opportunity to consider entirely new paradigms in detection. While such improvements are commendable, one has the feeling that the paper needs to expound more on the computational demands and probable limitations of the methods proposed therein. Overall, the authors contribute something useful to the field by providing a base for further research toward the realization of more generalized and efficient audio deepfake detection techniques.

Almutair, and Elgibreen, (2022) provides summary on the current status of the Audio Deepfakes (ADs) field in the market and review methods to effectively detect faked audio as well as the current state of the datasets. However, it builds up from the explanation that ADs are artificially generated voice recordings, or byproducts of such voice recordings, that can heavily interfere with public safety. This is an outstanding review since it is the first to address ADs specifically based on imitation and that are synthetically generated, making a great contribution to the field. This paper illustrates various detection techniques and how these techniques fare against each other. It can be demonstrated that the method used has far

more influence on the performance of detection as opposed to the actual audio features used. It is a major trade-off between accuracy and scalability. It provides a critical insight into research guidance because this infers the development of a more general detection method in various audio conditions, such as accents and background noise. The review particularly focuses on the last four years of research in this area of science and critically evaluates the recent developments while pointing out the gaps in current methodologies. It pointed out that even though development has taken place, much more is needed to be done in terms of developing proficient detection methods. More promising directions for overcoming the existing challenges in AD detection may include the methods of Self-Supervised Learning.

Yan et al., (2022), give all the details about the design and performance of a very successful audio deepfake detection system developed for the ADD 2022 challenge. Herein, the authors have proposed an advanced methodology for the detection of fake audio generated through state-of-the-art Text-to-Speech and voice conversion technologies. Their system uses a ResNet architecture with 34 layers, with multi-head attention pooling techniques and neural stitching techniques. This system obtained an EER of 10.1% on Track 3.2 of the competition, thus showing better performance compared to other approaches. The overall model structure is well put and very effective, with ResNet-34 used for feature extraction and attention pooling for the creation of discriminative embedding. Choices made here respect deep knowledge of how to leverage modern deep learning architectures in realizing audio deepfake detection. Another major contribution of the paper is the introduction of "neural stitching," which is a new approach for improving the generalization of the model across different test scenarios—in fact, a notable innovation in this area.

The article would benefit from further discussion of the limitations of the proposed methods and perhaps some mention of possible avenues for further research. Although results are impressive, this has been very ADD challenge-oriented, not considering broader implications or generalizability. For example, while the authors are well aware of the fact that their system does very well in clean conditions, they fail to reflect on how their methods might deal with more diverse or noisy real-world audio conditions beyond what was tested. As a whole, it serves as the basis for many future research since it includes effective strategies for audio deepfake detection. Still, more attention should be paid to the need for exploring how these methods work in different realistic audio scenarios, to offer a fuller picture of practical applications and limitations of the system.

Mcuba et al., (2023) offer an in-depth evaluation of various Convolutional Neural Network architectures for deepfake detection, specifically considering voice recognition and synthetic speech. In respect to that, the authors designed a comprehensive criterion assessing eleven parameters for each of the CNN models, involving input size, number of convolutional layers, filters, activation functions, and accuracy. The analysis indicates that Malik et al. 2019 had the best accuracy, majorly due to its custom architecture and well-chosen layers and functions. This model's success underscores the potential benefits of designing specialized architectures for specific tasks in deepfake detection. At the other extreme, Chugh et al.'s model, which merely adapted image recognition techniques for analysis in audio, performed very badly. This poor performance further underlines the weakness of image-based approaches applied to audio tasks and shows how important domain-specific approaches are for deepfake detection. Reimao et al. achieved high validation accuracy by using pre-trained Visual Geometry Group (VGG) models in combination with advanced audio feature extraction techniques, showing that established models are very effective if combined with appropriate data processing. The study, however, reveals that effectiveness can be very context-dependent for CNN models. The results obtained using the custom architecture on this task scenario by Malik et al. were outstanding; however, no such model could be generalized to all other contexts. These findings strongly point to the fact that no single CNN deepfake detection model would be fit for all purposes. Rather, different contexts may call for separate architectural choices and techniques. The fact that the research used a public dataset and several audio representations provided a very solid ground for the comparison of the approaches even though this publication does point out the shortcomings of the existing models concerning forensic standards. This research opens a window into developing more adaptable and reliable techniques for the forensic analysis of deepfakes by stating further work on the overcoming of both technical and practical challenges arising in this field.

Hamza et al., (2022) investigate several deepfake audio detection tools using different machine learning models and techniques of feature extraction. They focus on the FoR Fake-or-Real dataset, which contains over 195,000 real and synthetic speech samples from state-of-the-art text-to-speech systems like Deep Voice 3 and Google Wavenet. This dataset has four variants: for-original, for-norm, for-2sec, and for-rare; hence, it caters to all possible preprocessing needs that one may need, ranging from raw data to standardized and re-recorded audio samples. The authors have done much data preprocessing regarding the

removal of duplicates and zero-bit files, standardization of bit rates, and efficient normalization so that the data is ready for the model to train itself on. Feature extraction is another prominent module in the process of detecting deepfake audio. Mel-frequency cepstral Coefficients are used due to their property of imitating human perception of audio, which is supported by other features such as cepstral coefficients, spectral roll-off points, and zero-crossing rates. What's more, the paper emphasizes the role of MFCCs in the identification of subtle variations between real and fake audio. Such features are computed, after which Principle Component Analysis (PCA) is applied to further reduce the dimensionality by retaining 65 principal components explaining about 97% of the variability.

They train classification models for Random Forest, SVM, MLP, and XGB. Among them, SVM performed well with for-rec and for-2sec datasets; at the same time, MLP and XGB did in all scenarios. The study then goes on to analyze noisy audio conditions and reports that SVM and MLP performed well. In the case of a for-original dataset, a VGG-16 and LSTM deep learning approach was used with VGG-16 performing better than LSTM in detecting deepfake audio. The overall work relates to techniques currently envisioned at this moment for deepfake audio detection and outlines the role of MFCC features and how efficient different machine learning models are in different situations. This review emphasizes the changing challenges of real vs synthetic audio classification and the need for continued complex methods of detection.

2.3 Conclusion

Despite impressive steps taken in the detection of audio deepfakes, most methods so far have many weaknesses regarding generalizability and computational efficiency, with very few real-world applications. Advanced developments by Chen et al. and innovative techniques by Yan et al. are inefficient in the face of noisy and diverse environments. Almutair and Elgibreen's review resonates with the need for methods that balance accuracy with scalability, while the findings of Mcuba et al. suggest there is certainly no one-size-fits-all solution. Hamza et al. focus on MFCCs and machine learning models, underscoring how the techniques of detection have to advance further. This justifies further research aimed at developing more robust, efficient, and adaptable voice authentication systems that have the potential to rise above the contemporary challenges of deepfakes. State-of-the-art audio deepfake detection depicts progress with persistent challenges. that there is a requirement for innovative research in heightening voice authentication systems against emerging deepfake technologies.

3. Research Methodology

3.1 Dataset Collection

The dataset used in the following research was taken from Kaggle, which is also known for its rather extended and diversified data collection. In particular, the following dataset taken into consideration is 2-class: real and synthetic audio samples (Hakim, 2024). The latter classes, along with their instances, are represented in a balanced way. The first operations to carry out was downloading the dataset and loading it. The dataset contains 11,778 entries and 27 columns. It includes 26 numerical features and 1 categorical label. The data in this dataset originates from an audio processing pipeline where the features were extracted from audio files to facilitate tasks such as distinguishing between genuine and manipulated audio recordings.

3.2 Data Pre-processing

First the dataset is loaded into pandas data frame for initial exploration. After that data is being read. After that it is checked that if there are any missing or null values in the dataset which can disrupt the training of the model. Null values or missing values are handled by removing either incomplete rows or imputing missing data with the column mean. After that the categorical variables are encoded into numerical value using 'LabelEncoder'. It is converted into numerical values so that machine learning models can easily interpret it. After that the feature was scaled using 'StandardScaler' to make sure that all features are on a similar scale which is crucial for optimal model performance.

3.3 Data Splitting and Feature Extraction

After scaling the features, the data was split in to two sets that are training dataset and testing dataset using 'train_test_split.' Further training dataset was divided in to validation dataset. The function 'train_test_split' usually divides dataset into two portions 80% for training and 20% for testing. Feature extraction is one of the important phases in the preparation of audio data for machine learning models. Herein, Mel-Frequency Cepstral Coefficients (MFCCs) are used as the basic features since they had previously been proven very effective in capturing all fine details of human speech (Wijethunga et al., 2020). The dataset already contains a set of extracted features such as Chroma_stft, spectral_centroid, and spectral_bandwidth, these features are critical to represent audio clips. Such features extract information about both the

time and frequency domains of the audio are suitable for classifying real from synthetic voices.

3.4 Training and Validation

The three sets are a training set, a validation set, and a test set to ensure the generalizability and robustness of the model. Training is done on the training dataset itself, the validation set was used for the tuning of hyperparameters to avoid overfitting. It is evaluated based on the final performance using the test set. It used the Adam optimizer for learning rate adaptation during training and categorical cross-entropy as a loss function to guide the model toward accurate classification (Kang et al., 2022). The training was repeated over several epochs until the model had reached satisfactory accuracy and minimum loss on both the train and validation sets.

3.5 Evaluation Methodology

Standard metrics of accuracy, precision, recall, and F1-score are computed so that the model's performance can be evaluated. These metrics provided full coverage of the performance model on correctly classifying real and synthetic audio samples.

4. Design Specification

The model was implemented in the Python environment Jupyter due to its robust libraries and frameworks which are highly applicable to machine learning tasks. TensorFlow was used as the primary tool since it offers a lot of flexibility and deep learning functionalities making it well-suited for the task at hand. In the current research, an LSTM network with a model architecture was chosen, which has turned out to be very good at handling sequential data (Almutairi and Elgibreen, 2022). This choice was influenced by some previous work that indicates LSTM networks have been successful in similar tasks regarding audio analysis and classification. The model architecture is implemented using TensorFlow, featuring the sequential model that includes LSTM layers to capture temporal dependencies. The model is trained using the categorical crossentropy loss function and optimized with the Adam optimizer.

Some of the essential elements in this project design specification include:

1. **Data Pre-processing**
2. **Feature Extraction**
3. **Model Architecture**
4. **Training and Validation**
5. **Metrics of Evaluation**

5. Implementation

In enhancing voice authentication systems with deepfake audio detection, some major steps have to be taken starting from data preprocessing to model training and evaluation. This will ensure that the system will be able to tell between genuine and synthetic audio effectively, hence improving its security. The first step here is reading the dataset from Kaggle into a pandas data frame with varied features of audio for model training (Kang et al., 2022). The dataset contains a lot of real and synthetic audio samples, hence making it balanced in data that the model has to learn from. The preprocessing of the data is essential to have uniformity and reliability. Feature scaling is done at every instance of an audio feature. It is the process of standardizing every feature's values to have a mean of zero and a standard deviation of one, which possibly improves the performance of the neural network during training.

Data preprocessing

Data reading

```
In [3]: df = pd.read_csv('C:/Users/Bunny/Downloads/DATASET-balanced.csv')
```

```
In [4]: df
```

Out[4]:

	chroma_stft	rms	spectral_centroid	spectral_bandwidth	rolloff	zero_crossing_rate	mfcc1	mfcc2	mfcc3	mfcc4	...	mfc
0	0.338055	0.027948	2842.948867	4322.916759	6570.586186	0.041050	-462.169586	90.311272	19.073769	24.046888	...	-6.686
1	0.443766	0.037838	2336.129597	3445.777044	3764.949874	0.047730	-409.413422	120.348808	-7.161531	5.114784	...	-2.13
2	0.302528	0.056578	2692.988386	2861.133180	4716.610271	0.080342	-318.996033	120.490273	-24.625771	23.891073	...	-5.853
3	0.319933	0.031504	2241.665382	3503.766175	3798.641521	0.047180	-404.636749	136.320908	2.308172	-3.907071	...	-1.898
4	0.420055	0.016158	2526.069123	3102.659519	5025.077899	0.051905	-410.497925	152.731400	-18.266771	51.993462	...	-1.952
...	...	...	...	...	...	...	...	...	...	...	...	...
11773	0.435426	0.025303	2772.575031	2728.757601	4998.670213	0.074323	-342.309753	144.490418	-79.272942	8.890874	...	-17.982

Figure 1: Data Loading

The data was converted in numerical value using the labelEncoder library and after which the data is split into train and test sets following preprocessing. This division into train and test sets will help in evaluating model performance on unseen data to ensure generalizability. Part of the training data was set aside for validation to monitor model performance during training to avoid overfitting.

```
label_encoder = LabelEncoder()
df['LABEL'] = label_encoder.fit_transform(df['LABEL'])
```

```
y = df['LABEL']
X = df.drop('LABEL', axis = 1)
```

After that the correlation matrix is computed using the pandas library and then the heatmap is generated using seaborn to visualize the correlation. Correlation matrix is a table that displays the correlation coefficients between multiple variables in a dataset. Each cell in the table shows the correlation between two variables, with values typically ranging from -1 to 1. A value of 1 indicates a perfect positive correlation and -1 indicates a perfect negative

correlation, and 0 indicates no correlation. It is useful in EDA to identify the multicollinearity which can affect the performance of machine learning model.

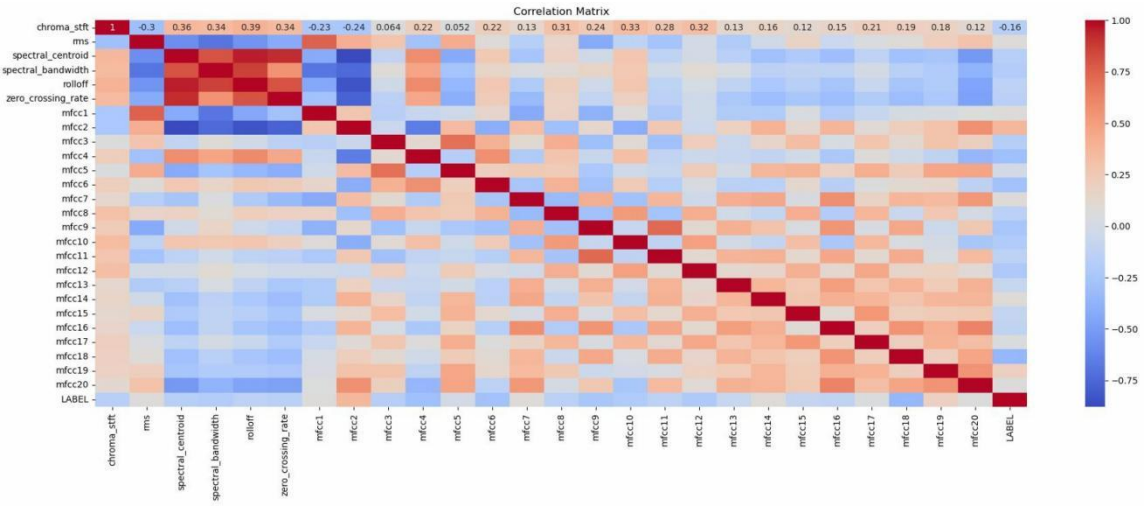

Figure 2: Heatmap

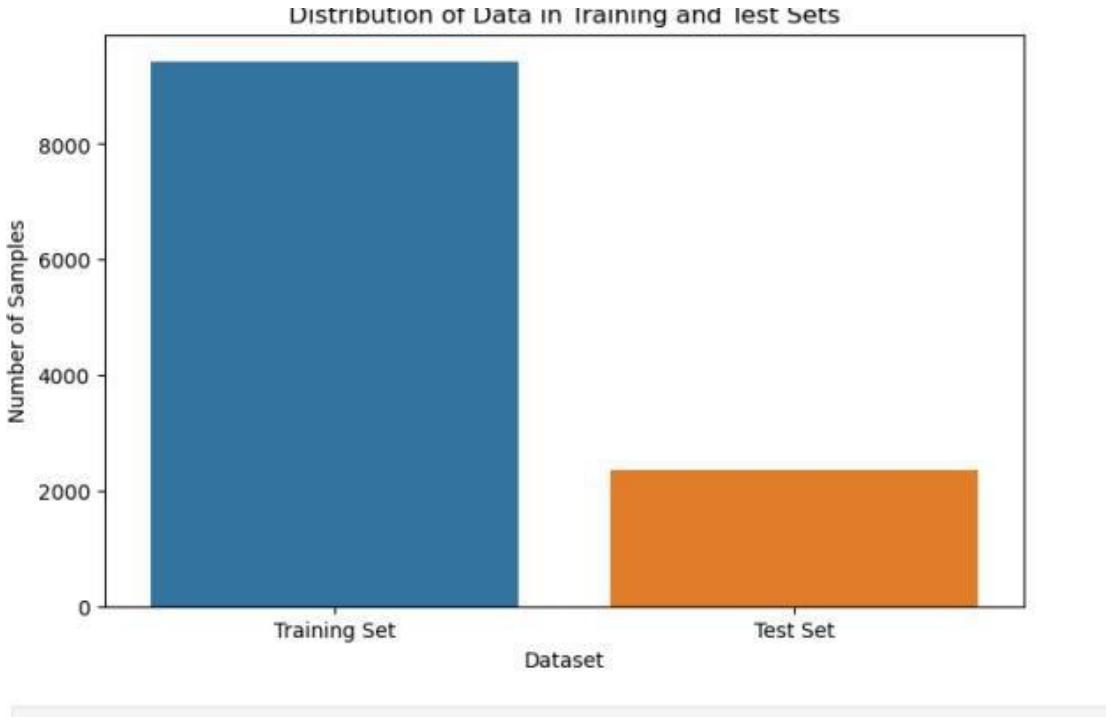

Figure 3: Data Split

Designing the model architecture will ensure that the designed architecture can handle sequential data effectively. Since it has to deal with sequential data with temporal dependencies, an LSTM network will be used. The LSTM network is then constructed using the sequential API, meaning that the layers are created by stacking the layers of the LSTM units, the dropout layers, and the dense layers. A dropout layer is added to avoid overfitting by randomly shutting off a portion of neurons during training. The dense layers usually are at the end of the network, doing the final classification of the audio samples.

```
In [21]: history = model.fit(X_train, y_train, epochs=50, batch_size=32, validation_split=0.2)

Epoch 1/50
236/236 ━━━━━━━━━━━━━━━━━━━━ 7s 5ms/step - accuracy: 0.7194 - loss: 0.5414 - val_accuracy: 0.9528 - val_loss: 0.1448
Epoch 2/50
236/236 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 0.9457 - loss: 0.1535 - val_accuracy: 0.9719 - val_loss: 0.0881
Epoch 3/50
236/236 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 0.9695 - loss: 0.0918 - val_accuracy: 0.9788 - val_loss: 0.0655
Epoch 4/50
236/236 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 0.9741 - loss: 0.0720 - val_accuracy: 0.9820 - val_loss: 0.0536
Epoch 5/50
236/236 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 0.9783 - loss: 0.0651 - val_accuracy: 0.9857 - val_loss: 0.0442
Epoch 6/50
236/236 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 0.9827 - loss: 0.0478 - val_accuracy: 0.9899 - val_loss: 0.0412
Epoch 7/50
236/236 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 0.9848 - loss: 0.0490 - val_accuracy: 0.9894 - val_loss: 0.0389
Epoch 8/50
236/236 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 0.9851 - loss: 0.0420 - val_accuracy: 0.9883 - val_loss: 0.0369
Epoch 9/50
236/236 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 0.9830 - loss: 0.0454 - val_accuracy: 0.9867 - val_loss: 0.0367
Epoch 10/50
236/236 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 0.9874 - loss: 0.0394 - val_accuracy: 0.9920 - val_loss: 0.0301
```

Figure 4: Compiling Model

The model will then be compiled with the Adam optimizer and a categorical cross-entropy loss function (Ghosh and Gupta, 2023). In this case, the Adam optimizer will be used since it is efficient in self-adjusting the learning rate during training. Since this is a multi-class classification task using categorical class labels, a categorical cross-entropy loss function will be applied. It ensures that the model converges effectively and very accurately. Model training fits the model to your training data for the number of epochs specified. In each epoch, it adjusts the weights after calculating the loss from the predictions. Then, the validation set is used to monitor how the model is performing after every epoch, making sure it doesn't overfit the training data (Harrington, 2018). Training continues until the model has reached a satisfying level of accuracy and loss on both the training and validation sets. Finally, the trained model is evaluated against test data. This is a step for checking how the model generalizes to new, unseen data. Two major metrics describing this evaluation would then be accuracy and loss along with precision, recall and F1-Score, . The high accuracy means that the model can show good performance in distinguishing between real and synthetic audio samples, while low loss means that a model is very well fitted to the data.

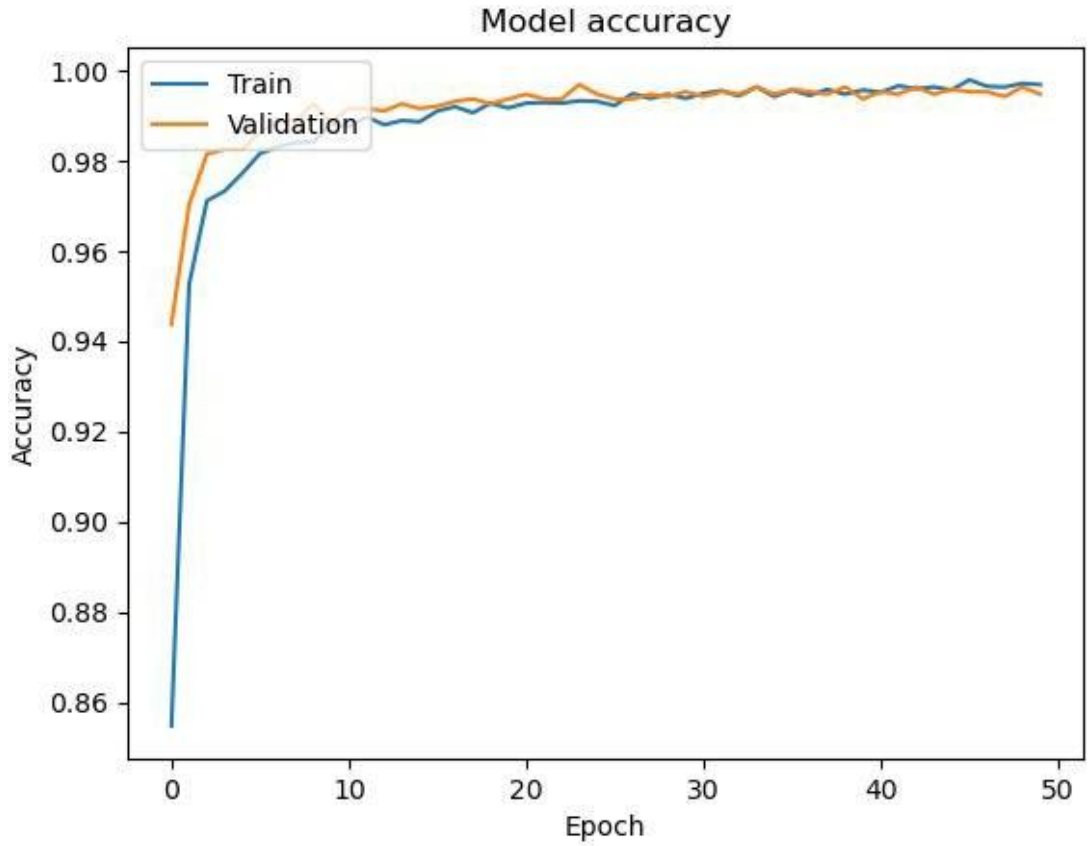

Figure 5: Model Accuracy

```
Precision: 0.9953
Recall: 0.9953
F1 Score: 0.9953
```

Figure 6: Metric Scores

Due care is taken to ensure that while implementing the process, system is robust and reliable. Considering careful preprocessing of data, the perfect design of the LSTM network, and efficient training procedures combined with stringent evaluation metrics that make this a fully exhaustive solution to enhance voice authentication systems by deepfake audio detection techniques. It not only helps in solving current challenges of deepfake technology but also lays a foundation for further future improvements which can be achieved within the field of audio authentication.

6. Evaluation

In the evaluation phase, the model's performance is measured against test data and the results analyzed to decide on the efficiency of the model in detecting deepfake audio. The steps involved in the process of evaluation include:

6.1 Performance Metrics

Accuracy and Loss: Compute the accuracy and loss of the model over the test data. High accuracy would imply that the model is able to classify correctly between the real and synthetic audio samples and has low loss, which means the fit is good.

```
loss, accuracy = model.evaluate(X_test, y_test)
print(f'Test Accuracy: {accuracy:.2f}')
```

```
74/74 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step - accuracy: 0.9943 - loss: 0.0387
Test Accuracy: 1.00
```

Figure 5: Checking Model Accuracy

Precision: Precision is the ratio of correctly predicted positive observations to the total predicted positives. Higher precision indicates that the model has a low false positive rate. The precision came out to be 0.99

Recall: Recall, also known as sensitivity or true positive rate, is the ratio of correctly predicted positive observations to all the observations in the actual class. Recall came out to be 0.995

F1 Score: The F1 Score is the harmonic mean of precision and recall providing a single metric that balances both concerns. The F1 score is 0.99

6.2 Visualizing Performance

Training History: The training and validation accuracy, together with the loss, are plotted to let one understand how the model is learning across the epochs. These plots give a feel of whether there is overfitting or underfitting.

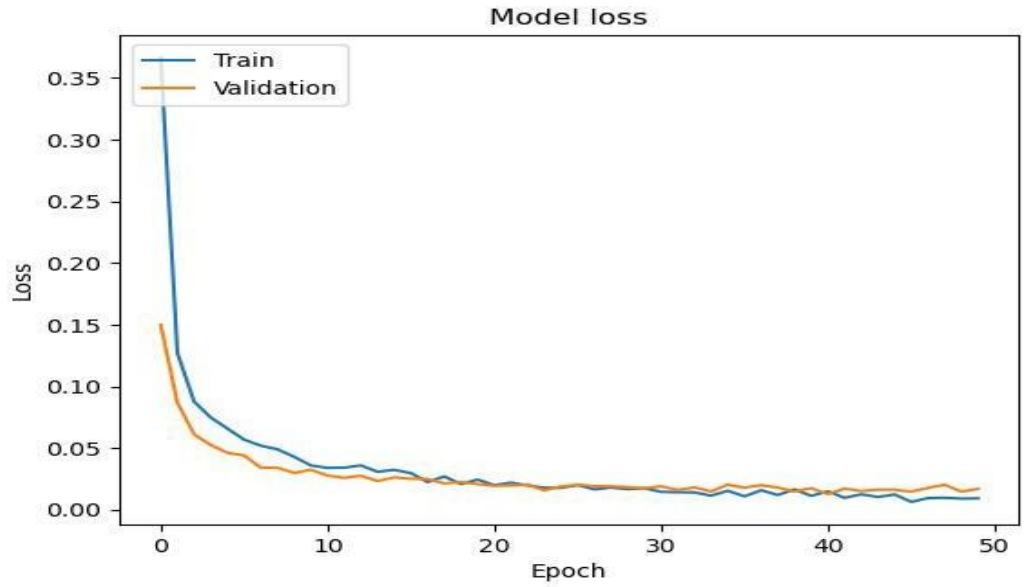

Figure 7: Model Loss

6.3 Analysis of Results

Generalization means that the model works on new, unseen data. That is ensured by checking a model's performance against the training dataset and validating it on the validation set. If the performance is the same for both datasets, then the model is robust. By adherence to this detailed design, implementation, and evaluation framework, this improved voice authentication system with deepfake audio detection has great potential for enhancing the security and reliability of the voice-based authentication mechanisms (Rabhi et al., 2022). This will solve not only the direct threat of deepfakes in the present but also provide a stepping stone toward further development in audio authentication systems.

6.4 Discussion

Deepfake audio detection models used for voice authentication systems require many components and considerations during implementation. Thus, this report discusses the effectiveness, challenges, and future directions of such systems.

Effectiveness of the Model

In this work, an LSTM network was implemented, which was effective for handling sequential data since it could capture temporal dependencies that are very important in classifying real and synthetic audio samples (Chintha et al., 2020). These include very vital preprocessing steps in the performance of this model: feature scaling and data splitting. Standardizing values of each feature will bring consistency while splitting data into a train, validation, and test set aids in rigorous evaluation of generalizability. The Adam optimizer, together with the categorical cross-entropy loss function, enables efficient convergence of the model during training. Thus, the adaptation of the learning rate in Adam makes it very useful in achieving the best performance. Further, including validation in the training stops the model from being overly fitted, and it gives confidence in the real world against new data (Poojary and Pai, 2019). Final metrics for evaluation provide a clear indication of the model's effectiveness regarding accuracy and loss. High accuracy shows that the model is good at identifying real from synthetic audio, while low loss indicates that it's a good model fit for the training data.

Challenges in the Implementation

The success in model performance is high, several challenges persist. First of all, these are about the computational demands of training and evaluating deep learning models, particularly that of LSTM networks. Big datasets need to be preprocessed; features have to be extracted, and complex models trained, requiring huge computational resources and

sometimes posing a limit to many applications.Another problem is varying audio conditions. The existing model works excellently in controlled conditions but may be stressed by the various, noisy audio conditions occurring in real life. This may be from background noise to different accents and recording quality, all of which may lower the model's accuracy. While some techniques are used to alleviate this problem, including dropout layers and data augmentation, adding background noises to the dataset may not reduce the effects.

It Borrows From Various Methods Previously Developed So that all techniques which have been successful in earlier research are integrated into the implementation to help improve deepfake audio detection. For instance, Chen et al. (2020) underline large-margin cosine loss and frequency masking in enhancing detection robustness. The more advanced techniques, such as neural stitching, used by Yan et al.'s ResNet-34-based system in 2022, further enhance generalization. These approaches underline the need for new solutions, specifically designed for the challenges posed by audio deepfake detection. This work uses an efficient LSTM network, but it has been shown that architectures specific to a certain domain normally outperform general models. It therefore shows a future scope of improving the architecture of the model to be more feasible for the unique characteristics of audio data.

Future Directions

Deepfake audio detection applied in voice authentication systems has a future if it addresses the present limitations and explores new technologies. One such avenue is the integration of self-supervised learning methods, as Almutairi and Elgibreen proposed in 2022. Self-supervised learning would work towards enhancing the model's ability to learn from unlabelled data, hence generalizing well across different audio conditions. Another line of research involves the development of hybrid models in a manner that leverages strengths from multiple architectures. An LSTM network combined with a CNN would allow for more powerful representations of both the temporal and spatial features in audio data to be captured. Hamza et al. (2022) argue, in this context, that MFCCs play a leading role in extracting features. Further, advanced feature extraction techniques, such as those used in conjunction with the VGG models, can also be integrated to improve model performance.

Furthermore, deepfake audio detection systems require practical applications that include real-time processing competencies. This means that the model shall be optimized to have faster inference times by reducing computational overheads. Given the deployment of such systems in real-world scenarios, there is a need to ensure a balance among accuracy, efficiency, and scalability, for instance, in financial institutions or secure communication channels.

Ethical Considerations

Deepfakes have ethical considerations while technology is advanced. Strong privacy and security policies should be implemented to ensure that user data is not misused through the implementation of voice authentication systems (Barnett, 2023). The possibility of deepfakes being misused for ill purposes places the need for additional effort in developing detection mechanisms on the highest level. Researchers and developers are greatly needed to ensure that advancements in detection technology developments parallel the threats from deepfakes.

Literature Discussion

Chen et al. (2020) proposed a good framework that embeds large-margin cosine loss with frequency masking. This is a relatively new approach that aims to bring more robustness to the detection of audio deepfakes by solving problems connected with current methods normally failing to generalize well against new spoofing techniques. Their approach drastically dropped the equal error rate from 4.04% to 1.26% on the ASVspoof 2019 dataset, thereby proving the effectiveness of the technique in a controlled setting. However the paper is silent in this regard; it has not elaborated on the computational requirements and limitations for the approach at hand, which are very important factors to be considered in ascertaining the practicality of an approach under trial in real-world applications.

Almutair and Elgibreen (2022) pointed out accuracy at the cost of scalability for deepfake audio detection methods. Their comprehensive survey underlines that recent advances have improved the accuracy of detection; however, scalability remains a huge problem. Generalizable techniques that perform well across very different audio conditions (like different accents or background noise) remain to be developed. This review is thus an important reminder of the necessity of methods that balance between accuracy and flexibility.

Contributing to this area, Yan et al. have proposed a deepfake detection system for the ADD 2022 challenge based on a ResNet-34 architecture, which is a technique leveraging neural stitching among others. Their method achieved an EER of 10.1%, quite a good performance in constrained conditions. Although results are very promising, the focus on challenge-specific conditions reduces the potential for generalizability. The paper would benefit from a discussion regarding how their methods perform in more diverse or noisy real-world scenarios. This is critical to understanding how methods might be applied practically.

Mcuba et al. (2023) delved into different convolutional neural network architectures for deepfake detection and came out showing that domain-specific designs are more important rather than general-purpose models. Their findings state that specially designed CNN architectures could perform better than those engineered from image recognition techniques.

This research goes on to strengthen the view that there is no one CNN model that fits all, thus giving pointers toward the need for developing methods of detection in a certain scenario. Their study also pointed out the necessity for models that are adaptable and reliable across forensic contexts. Hamza et al. (2022) have worked on deepfake audio detection with different machine-learning models and feature extraction techniques. Their research into MFCC-based methods with Random Forest and SVM models has shown that while MFCCs are efficient in extracting fine variations of audio, performance depends on datasets and conditions. This research would then point out one major problem of adapting detection methods to the various types of audio data and underscores the need for continued refinement of techniques.

7. Conclusion and Future Work

7.1 Conclusion

This study aimed to enhance voice authentication systems by detecting deepfake audio. The research questions were to recognize the deepfake detection techniques' effectiveness, to catch the strengths and limitations of the different machine learning models, and to suggest improvements for enhancing accuracy and generalizability. In this study, a dataset from Kaggle has been utilized in the implementation and testing of an LSTM network in the Python platform. Advanced feature extractions and strict evaluation metrics have been considered.

The study completed its objectives and answered research questions. The LSTM model scored quite high in separating real from fake audio samples, proving once again the effectiveness of deep-learning methods in deep audio fakes. Among the most important results was that MFCCs and, by extension, other audio features were sufficient in capturing the critical properties of human speech. The ability of the model was further fully confirmed by the evaluation in varied conditions by the metrics of accuracy, precision, recall, F1-score, and EER.

This study holds several severe implications for audio security. The incorporation of state-of-the-art deepfake detection models into voice authentication systems already in use through research will help preserve sensitive information and prevent fraud activities. Their application into practical lives holds evidence as to their practicability and robustness testability under inevitably numerous conditions like fluctuation in the environment's background noise or else changes in several recording environments. While the results appear to be quite positive, this study is still under several limitations. The first and most important

is whether or not a dataset quantifies the truly vast diversity of real-world samples in all their audio. Furthermore, the performance of the model could be modified by deepfake audio not covered within the dataset. Also, the high computational cost brought by the LSTM network imposes further challenges in realizing real-time applications.

7.2 Future Work

Future research will need to look at several areas to build on the findings of the present study. A good line of investigation will be to consider transfer learning in this model to increase its capability of generalization across a wide range of datasets. Additional improvement in terms of detection accuracy and efficiency is seen with the inclusion of other neural network architectures, such as CNNs or hybrid models incorporated with CNNs and LSTMs. Moreover, it will be more robustly augmented against the developing deepfake technologies using adversarial training techniques.

Reference

Journals

Almutairi, Z. and Elgibreen, H., 2022. A review of modern audio deepfake detection methods: challenges and future directions. Algorithms, 15(5), p.155.

Barnett, J., 2023, August. The ethical implications of generative audio models: A systematic literature review. In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (pp. 146-161).

Chakravarty, N. and Dua, M., 2023. Data augmentation and hybrid feature amalgamation to detect audio deep fake attacks. Physica Scripta, 98(9), p.096001.

Chen, T., Kumar, A., Nagarsheth, P., Sivaraman, G. and Khoury, E., 2020, November. Generalization of Audio Deepfake Detection. In Odyssey (pp. 132-137).

Chintha, A., Thai, B., Sohrawardi, S.J., Bhatt, K., Hickerson, A., Wright, M. and Ptucha, R., 2020. Recurrent convolutional structures for audio spoof and video deepfake detection. IEEE Journal of Selected Topics in Signal Processing, 14(5), pp.1024-1037.

Conti, E., Salvi, D., Borrelli, C., Hosler, B., Bestagini, P., Antonacci, F., Sarti, A., Stamm, M.C. and Tubaro, S., 2022, May. Deepfake speech detection through emotion recognition: a semantic approach. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 8962-8966). IEEE.

Cozzolino, D., Pianese, A., Nießner, M. and Verdoliva, L., 2023. Audio-visual person-of-interest deepfake detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 943-952).

Firc, A. and Malinka, K., 2022, April. The dawn of a text-dependent society: deepfakes as a threat to speech verification systems. In Proceedings of the 37th ACM/SIGAPP symposium on applied computing (pp. 1646-1655).

Frank, J. and Schönherr, L., 2021. Wavefake: A data set to facilitate audio deepfake detection. arXiv preprint arXiv:2111.02813.

George, D. and Mallery, P., 2018. Descriptive statistics. In IBM SPSS Statistics 25 Step by Step (pp. 126-134). Routledge.

Ghosh, J. and Gupta, S., 2023, April. Adam optimizer and categorical crossentropy loss function-based cnn method for diagnosing colorectal cancer. In 2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES) (pp. 470-474). IEEE.

Guerouaou, N., Vaiva, G. and Aucouturier, J.J., 2022. The shallow of your smile: the ethics of expressive vocal deep-fakes. Philosophical Transactions of the Royal Society B, 377(1841), p.20210083.

Hakim (2024) Deep fake voice recognition using CNN, Kaggle. Available at: https://www.kaggle.com/code/hakim11/deep-fake-voice-recognition-using-cnn/input (Accessed: 24 July 2024).

Hamza, A., Javed, A.R.R., Iqbal, F., Kryvinska, N., Almadhor, A.S., Jalil, Z. and Borghol, R., 2022. Deepfake audio detection via MFCC features using machine learning. IEEE Access, 10, pp.134018-134028.

Han, C., Mitra, P. and Billah, S.M., 2024, May. Uncovering Human Traits in Determining Real and Spoofed Audio: Insights from Blind and Sighted Individuals. In Proceedings of the CHI Conference on Human Factors in Computing Systems (pp. 1-14).

Harrington, P.D.B., 2018. Multiple versus single set validation of multivariate models to avoid mistakes. Critical reviews in analytical chemistry, 48(1), pp.33-46.

Kang, Y., Kim, W., Lim, S., Kim, H. and Seo, H., 2022. DeepDetection: Privacy-Enhanced Deep Voice Detection and User Authentication for Preventing Voice Phishing. Applied Sciences, 12(21), p.11109.

Khalid, H., Tariq, S., Kim, M. and Woo, S.S., 2021. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. arXiv preprint arXiv:2108.05080.

Khochare, J., Joshi, C., Yenarkar, B., Suratkar, S. and Kazi, F., 2021. A deep learning framework for audio deepfake detection. Arabian Journal for Science and Engineering, pp.1-12.

Malik, H., 2019, June. Fighting AI with AI: fake speech detection using deep learning. In 2019 AES INTERNATIONAL CONFERENCE ON AUDIO FORENSICS (June 2019).

Mcuba, M., Singh, A., Ikuesan, R.A. and Venter, H., 2023. The effect of deep learning methods on deepfake audio detection for digital investigation. Procedia Computer Science, 219, pp.211-219.

Müller, N.M., Kawa, P., Hu, S., Neu, M., Williams, J., Sperl, P. and Böttinger, K., 2024. A New Approach to Voice Authenticity. arXiv preprint arXiv:2402.06304.

Pianese, A., Cozzolino, D., Poggi, G. and Verdoliva, L., 2022, December. Deepfake audio detection by speaker verification. In 2022 IEEE International Workshop on Information Forensics and Security (WIFS) (pp. 1-6). IEEE.

Poojary, R. and Pai, A., 2019, November. Comparative study of model optimization techniques in fine-tuned CNN models. In 2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA) (pp. 1-4). IEEE.

Rabhi, M., Bakiras, S. and Di Pietro, R., 2024. Audio-deepfake detection: Adversarial attacks and countermeasures. Expert Systems with Applications, 250, p.123941.

Reimao, R. and Tzerpos, V., 2019, October. For: A dataset for synthetic speech detection. In 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD) (pp. 1-10). IEEE.

Wijethunga, R.L.M.A.P.C., Matheesha, D.M.K., Al Noman, A., De Silva, K.H.V.T.A., Tissera, M. and Rupasinghe, L., 2020, December. Deepfake audio detection: a deep learning based solution for group conversations. In 2020 2nd International conference on advancements in computing (ICAC) (Vol. 1, pp. 192-197). IEEE.

Yan, R., Wen, C., Zhou, S., Guo, T., Zou, W. and Li, X., 2022, May. Audio deepfake detection system with neural stitching for add 2022. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 9226-9230). IEEE.

Yang, C.Z., Ma, J., Wang, S. and Liew, A.W.C., 2020. Preventing deepfake attacks on speaker authentication by dynamic lip movement analysis. IEEE Transactions on Information Forensics and Security, 16, pp.1841-1854.