Siamese Networks for Audio Spoofing Attack Detection

Submitted By

Shah Rutva Jignesh

20MCED12

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2022

Siamese Networks for Audio Spoofing Attack Detection

Major Project

Submitted in partial fulfillment of the requirements

for the degree of

Master of Technology in Computer Science and Engineering

Data Science

Submitted By

Shah Rutva Jignesh

(20MCED12)

Guided By

Dr Sapan H. Mankad

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2022

Certificate

This is to certify that the major project entitled **"Siamese Networks for Audio Spoofing Attack Detection "** submitted by **Shah Rutva Jignesh (20MCED12)**, towards the partial fulfillment of the requirements for the award of degree of Master of Technology in Computer Science and Engineering - Data Science of Nirma University, Ahmedabad, is the record of work carried out by her under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for examination. The results embodied in this Major Project Part-II, to the best of my knowledge, haven't been submitted to any other university or institution for award of any degree or diploma.

Dr Sapan H. Mankad
Internal Guide & Assistant Professor
CSE Department
Institute of Technology
Nirma University, Ahmedabad

Dr Gaurang Raval
Assoc. Prof. & PG Coordinator(M.Tech - DS)
CSE Department
Institute of Technology
Nirma University, Ahmedabad

Dr Madhuri Bhavsar
Professor & Head
CSE Department
Institute of Technology
Nirma University, Ahmedabad

Dr Rajesh N. Patel
Director
Institute of Technology
Nirma University, Ahmedabad

Statement of Originality

I, **Shah Rutva Jignesh**, **20MCED12**, give undertaking that the Major Project entitled **"Siamese Networks for Audio Spoofing Attack Detection "** submitted by me, towards the partial fulfillment of the requirements for the degree of Master of Technology in **Computer Science and Engineering - Data Science** of Institute of Technology, Nirma University, Ahmedabad, contains no material that has been awarded for any degree or diploma in any university or school in any territory to the best of my knowledge. It is the original work carried out by me and I give assurance that no attempt of plagiarism has been made.It contains no material that is previously published or written, except where reference has been made. I understand that in the event of any similarity found subsequently with any published work or any dissertation work elsewhere; it will result in severe disciplinary action.

Signature of Student
Date:
Place:

Endorsed by
Dr Sapan H. Mankad
(Signature of Guide)

Acknowledgements

It gives me immense pleasure in expressing thanks and profound gratitude to **Dr Sapan H. Mankad**, Assistant Professor, Computer Science and Engineering Department, Institute of Technology, Nirma University, Ahmedabad for his valuable guidance and continual encouragement throughout this work. The appreciation and continual support he has imparted has been a great motivation to me in reaching a higher goal. His guidance has triggered and nourished my intellectual maturity that I will benefit from, for a long time to come.

It gives me an immense pleasure to thank **Dr Madhuri Bhavsar**, Hon'ble Head of Computer Science And Engineering Department, Institute of Technology, Nirma University, Ahmedabad for her kind support and providing basic infrastructure and healthy research environment.

A special thank you is expressed wholeheartedly to **Dr Rajesh N. Patel**, Hon'ble Director, Institute of Technology, Nirma University, Ahmedabad for the unmentionable motivation he has extended throughout course of this work.

I would also thank the Institution, all faculty members of Computer Science and Engineering Department, Nirma University, Ahmedabad for their special attention and suggestions towards the project work.

- Shah Rutva Jignesh
20MCED12

Abstract

In recent times, audio spoofing attacks have become very common and serious in regard to Automatic Speech Verification(ASV) systems. There are many types of attacks possible like impersonation, speech synthesis, voice conversion or replay attack. Neural Networks mostly prove to perform well on such problems, and one of such is the Siamese Network which performs really well on simple data for classification. So in this paper we focused on detecting the audio spoofing attacks with the solution to it given by using Siamese Network. We have tried various different approach by combining different feature extraction techniques and various Siamese algorithms. We have used the popular Neural Network architecture, Siamese Networks which is hardly used in the audio domain. Audio data is difficult to work with, this method uses MFCC, Spectral Centroid, Spectrograms and Chroma Features for feature extraction, different python audio libraries like librosa, pyaudio, spafe etc. The result of this is combined with different Siamese networks and the performance is compared based on the Equal Error Rate(EER) of all these methods. So the method proposed in this paper is to use the very efficient Siamese Network algorithm for audio data and compare the performance of all variations and use it for detection of audio spoofing attacks.

Abbreviations

MFCC - Mel-frequency cepstral coefficients

CNN - Convolutional Neural Network

EER - Equal Error Rate

FAR - False Acceptance Rate

FRR - False Rejection Rate

ASV - Automatic Speaker Verification

GMM-UBM - Gaussian Mixture Model Universal Background Model

List of Figures

List of Tables

Contents

Chapter 1

Introduction

In this chapter we discuss about the Audio Spoofing attacks, its types and the ASVspoof 2017 dataset. The chapter discusses in brief about Audio spoofing attack and how Siamese Network is used for its classification process

1.1 Audio Spoofing Attack

Audio Spoofing Attack—Audio data is really hard to deal with, and so is to detect if the data is genuine or not. So for the attackers it becomes really convenient to not be caught and more difficult for the system to detect such attacks. Audio spoofing attack is being detected by many methods using different classification algorithms giving good accuracy. We are using the ASVspoof 2017 dataset. And the data of the fraud or spoofed class may be less and in that case the normal neural network may not give good performance. So we will be using the neural network architecture Siamese Network that is popular for precise prediction for smaller datasets.

1.1.1 Audio Spoofing Attack Defined

Spoofing refers to the disguise of communication or identity of the authorized person to fool someone with that data. Audio Spoofing attack is when the fraud is done using audio data with any medium. For the audio data we will be using the ASVspoof 2017 dataset which has genuine as well as spoofed class data. This can be classified by many factors of audio data such as, its speed, bandwidth, spectral features, background noise, depth the voice etc. For this we need the feature of the audio, which can be obtained by feature extraction methods like MFCC (Mel Frequency Cepstral Coefficients). Then these features are used for classification and prediction purpose. There are many types of audio spoofing attacks all different as mentioned below and shown in Figure 1.2

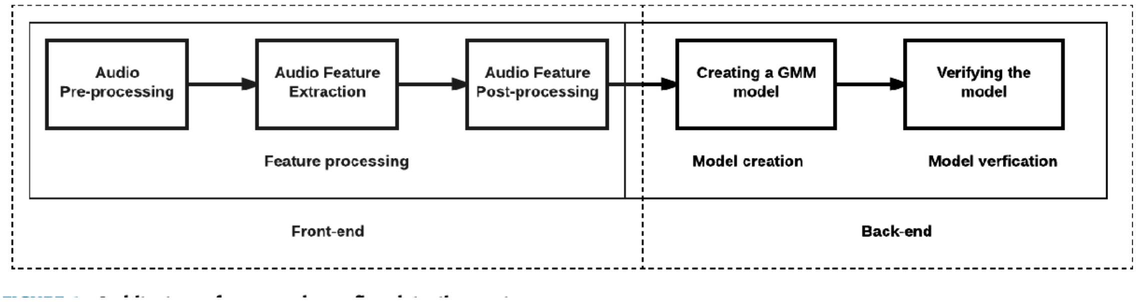
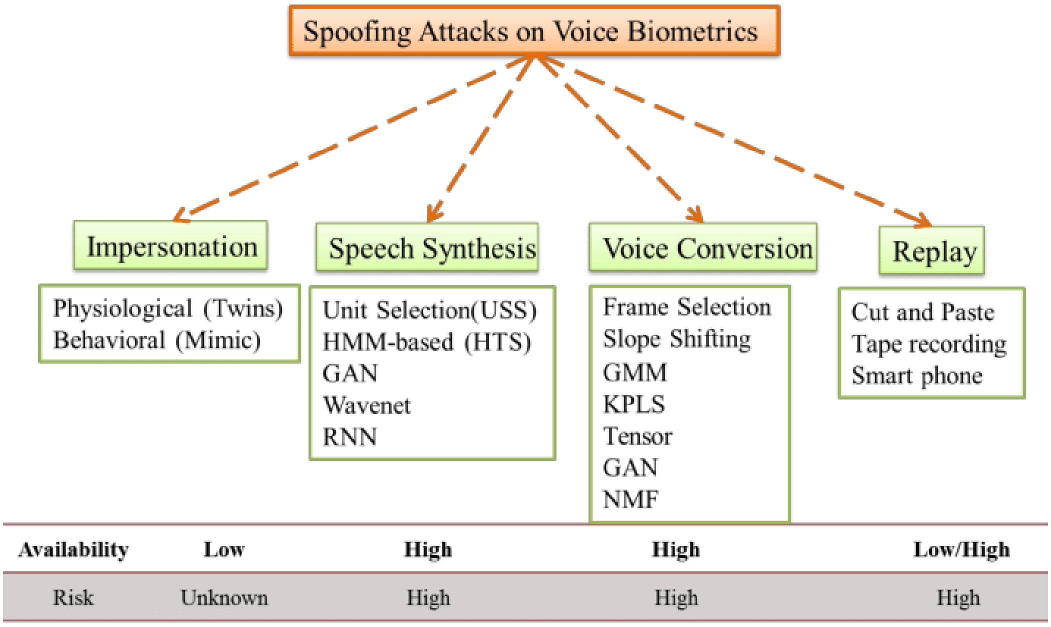


Figure 1.1: Audio Spoofing Attack

TYPES OF AUDIO SPOOFING ATTACKS:

- Impersonation Attack
- Speech Synthesis Attack
- Voice Conversion Attack
- Replay Attack



Availability	**Low**	**High**	**High**	**Low/High**
Risk	Unknown	High	High	High

Figure 1.2: Types of Audio Spoofing Attack

1.2 Neural Network Architecture: Siamese Networks

Bromley and LeCun initially proposed Siamese nets in the early 1990s to handle the challenge of signature verification as an image matching problem (Bromley et al., 1993). A Siamese neural network is a form of artificial neural network that computes similar output vectors from two separate input vectors using the same weights. Precompiling one of the output vectors creates a baseline against that the other output vector is assessed. Output of the Siamese network's goal is to distinguish between the two inputs X1 and X2. The network's output is a probability between 0 and 1, with a value closer to 0 indicating a prediction of dissimilar images and a value closer to 1 indicating a prediction of comparable images.

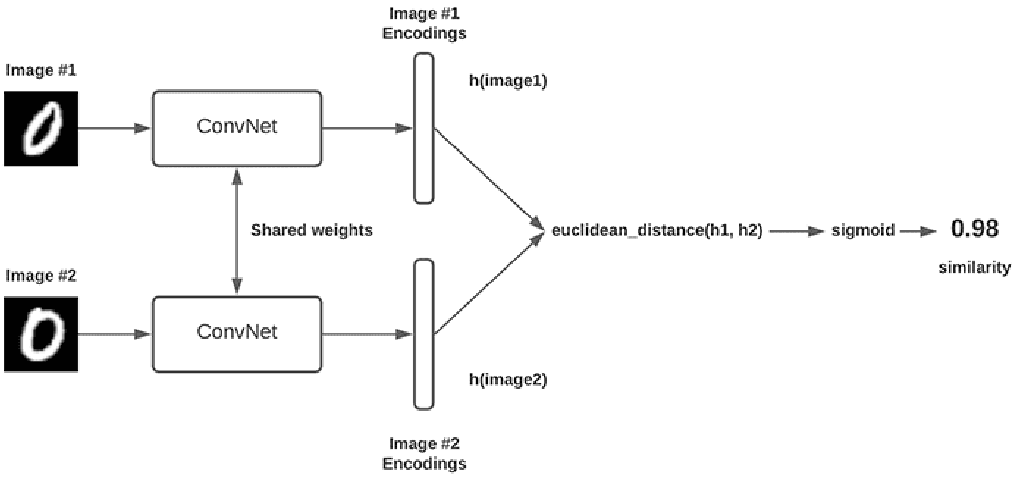


Figure 1.3: A basic siamese network architecture method takes two input photos (left), has identical CNN subnets for each input (mid), calculates the Euclidean distance between both the fully-connected layer outputs, and then uses the distance to determine similarity (right)

1.3 Siamese Network Implementation and Audio Spoofing Detection

Nowadays for every small or big problem, the solution can be obtained using Neural Networks. The enormous amount of data makes it even better, making the performance of the neural network more precise. But for the smaller available dataset, it may cause issues and result in degraded performance. And for the problems like fraud detection, and spoofing detection, the data for the fraud and spoofed class respectively is really less.

Face recognition and signature verification, for example, have extremely little data. For that, we have this different type of neural network with similar twin layers architecture, that is Siamese Networks. The Siamese Network requires very little data and yet gives a great accuracy prediction which makes it really popular

Chapter 2

Literature Survey

2.1 DEEP LEARNING IN AUDIO REPLAY DETECTION

These days spoofing detection is one of the needed research areas in the field of programmed speaker confirmation. The accomplishment of Programmed Speaker Verification Spoofing and Countermeasures (ASVspoof) Challenge 2015 affirmed the great perspective in recognition of unanticipated ridiculing preliminaries dependent on discourse blend and voice transformation procedures. Be that as it may, there is few investigates routed to replay parodying attacks that are bound to be utilized by non-proficient impersonators. This paper portrays the Speech Technology Center (STC) against parodying framework submitted for ASVspoof 2017 which is centered around replay assaults location. Here we investigate the effectiveness of a profound learning approach for arrangement of the referenced above task.

Trial results acquired on the Challenge corpora exhibit that the chosen approach outflanks present the status of the craftsmanship gauge frameworks in the wording of parodying identification quality. Their essential framework created an EER of 6.73% on the assessment part of the corpora which is 72% relative improvement over the ASVspoof 2017 pattern framework. In this paper, they investigated the relevance of the profound learning approach for arranging the issue of replay assault parodying detection. They examined single CNN and joined with RNN approaches. Their trials led on the ASVspoof 2017 dataset affirmed high effectiveness of profound learning structures for parodying recognition "in the wild". EER of the best person CNN framework was 7.34%. Our essential framework dependent on frameworks score combination gave 6.73% EER on the eval set[13].

2.2 COMPARISON OF FEATURE EXTRACTION METHODS

Automatic speaker Verification, similar to each other biometric framework, is helpless against ridiculing assaults. Utilizing a couple of moments of voice that is recorded of an authentic customer of a speaker confirmation framework, assailants can build up an assortment of mocking assaults that may deceive such frameworks. Identifying these assaults utilizing the sound signals present in the chronicles is a significant test. Most existing mocking location frameworks rely upon realizing the utilized mocking strategy. This investigation aims to overcome this limitation or drawback by examining hearty sound highlights, both traditional autoencoder-learned, which can be generalized to various types of replay spoofs. In addition, we include a detailed list of all the tools required to set up cutting-edge sound component position, pre-, and postprocessing, so that a (non-sound master) AI scientist can implement such frameworks. Then they evaluated the execution of our powerful replay parodying position system on 'in the wild' ASVspoof 2017 dataset using a wide variety and various blends of both extricated and machine trained sound highlights. A variety of new replay ridiculing arrangements are included in this dataset.

Their focus is on determining the features would ensure vigor, so they built their architecture around a standard Gaussian Mixture Model Universal Background Model (GMM-UBM). They purposefully investigate the general commitment of each list of capabilities at that stage. The combined models provide a tantamount exhibition with an EER of 12 in view of both realized sound highlights and machine learned includes separately. The model with best performance, with an 10.8 EER, is a half-breed architecture that incorporates both established and machine-learned features and is trained on a larger dataset, demonstrating the value of combining both types of features when constructing a robust ridiculing forecast model. They look at how a large number of sound highlights affect the presentation of a GMM-UBM based replay sound caricaturing discovery framework. One of the goals of this paper is to identify the key points to consider when15 building a strong model that works on a 'in the wild dataset,'[12] that is, without any prior knowledge of the pre-existing replay mocking technique. As well as altogether looking at the models fabricated on the various highlights, we likewise give an unmistakable technique for appropriate pre-and postprocessing of these highlights, which we expectation will be significant to different scientists. In our trials, we investigate both realized sound highlights, what's more, those educated by an autoencoder (i.e., utilizing a feed forward neural organization). The previous incorporates some accepted includes frequently utilized this field, just as some conceivably new highlights that would have the option to recognize veritable discourse from satirize one.

MFCCs, spectrogram, CQCCs, LPCCs, IMFCCs, RFCCs, LFCCs, SCFCs, SCMCs, and CCCs are among the highlights. An autoencoder also learns a new portrayal for each of these capabilities. To expand the preparation range, the autoencoder is used. As far as EER is concerned, any one of the models that have been used with these various capabilities has been presented. The subsequent exhibition is about 12 in terms of EER by using only recognized highlights or only auto encoder highlights. We achieve a common exhibition of 10.8 by constructing a mixture structure that incorporates both types of highlights. This contrasts with the current state of the art and stresses the importance of coordinating different types of sound highlights, both known and machine trained, in order to build a powerful model for replay spoofing recognition.

2.3 SIAMESE NETWORK FOR AUDIO SIMILARITY

In this paper the authors has worked on the similarity of two audio signals using Siamese LSTM Network. The selection of audio signal features and feature matching model is the essential technology in audio signal similarity detection. A method of using LSTM in the basic network section of the Siamese network is proposed to increase the accuracy of audio similarity estimation. First and foremost, they extracted the properties of the two audio signals' filter banks Then, to calculate the result, two feature matrices are sent into the network. Audio resemblance Experiments have shown that the Siamese LSTM is effective. Using FBank characteristics, a network can accurately detect similarity. consisting of two audio segments[8]. The experiment employed 60,000 audio segments separated into 30 categories as the dataset. Animal calls, human speech, percussive sounds from various materials, and so on are among the noises. A batch of 128 data pairs was used for training, with RMSprop as the optimizer. The loss on the verification set drops to 0.0715 after 150 rounds of training, and the accuracy rises to 0.9323. They employed five approaches for detecting acoustic similarity in the dataset for comparison.

- FBank + ordinary Siamese network
- MFCC+ ordinary Siamese network
- MFCC+Siamese LSTM network
- Fbank +Siamese LSTM network
- MFCC+Siamese LSTM network

From all the tried variants the accuracy proved to be the highest when FBank + Siamese LSTM Network were used.

2.4 ANALYZE EARLIER DETECTION METHODS

In this paper the author has proposed a method to classify the audio data into genuine and spoof using the SLIME algorithm. They analyzed the Convolutional neural network based method submitted at the ASVspoof 2017 challenge. They discovered that the classification of the audio more of depends on the first 400 milliseconds. Using this information and the slime algorithm they did experiments by interchanging the first and the last 400 milliseconds of the misclassified genuine and the correctly classified spoof with high confidence to get the end results which showed increase in the equal error rate for protecting the system and decrease in case of the attacker's perspective[10]. Thus the paper concluded that using LCNN(FFT) for classification and SLIME for generation of class explanation of temporal and spectral perspective

2.5 Literature Summary

Paper Title	**Year**	**Type**	**Method**	**Summary**
Audio Replay Spoof Attack Detection Using Segment based Hybrid Feature and DenseNet LSTM Network	2019	Paper	Hybrid feature extraction with Dense, LSTM and Dense-LSTM	In this paper the author proposes a method which is proven to be performing better 64.31 then the original method. The methods used for audio data feature extraction are MFCC, CQCC. But the proposed method uses a segment based hybrid method which takes hybrid features from both the methods that are MFCC and CQCC which are then trained using different methods, Dense, LSTM and hybrid architectures like Dense-LSTM. The proposed method performs really better then the baseline approaches in detection of audio replay spoof attack.
Robust Signal Classification Using Siamese Networks	2019	Paper	Siamese Convolutional Neural Network, Image Classification	The paper is about classifying images with noise that is, classifies even the similar wireless signal emitters across signal to noise ratio and a dataset of small size using Siamese Convolutional Neural Network and compared the performance with the basic convolutional neural network output[3]. The model was trained on compressed spectrogram images to differentiate out the randomized signals amongst the modulated signals. And the performance of the proposed model proved to be more efficient and improved classification.

Table 2.1: Literature summary

Paper Title	Year	Type	Method	Summary
Audio Feature Extraction Based on Sub Band Signal Correlations for Music Genre Classification	2018	Paper	Support Vector Machine	The authors provided a set of new low-level audio features which were based on correlations amongst subband audio signals decomposed using the undecimated wavelet transform. The experimental findings which were on the GTZAN dataset, resulted that the suggested method outperformed the standard methods with an accuracy of 81.5 percent under the assumption that Support Vector Machine is employed for classifier learning[7].
Deep Learning for Audio Signal Processing	2019	Paper	Deep Learning	This study gives a review of the state-of-the-art deep learning algorithms for audio signal processing, given the recent increase in deep learning breakthroughs[6].
Replay attack detection with raw audio waves and deep learning framework	2019	Paper	Convolutional Neural Network	The author in the paper has proposed a 1D ConvNet system that includes raw audio waves as one of its properties. This approach achieves an EER of 0.41 percent on the development set and 5.29 percent on the evaluation set, beating the best submission to the ASVspoof 2017 challenge, which had an EER of 3.95 percent on the development set and 6.73 percent on the evaluation set[11].

Table 2.2: Literature summary

Chapter 3

Audio Data and Audio Spoofing Attack

Audio signal and format—An audio signal is nothing but a sound reflection that is usually defined by an electrical voltage level for analogue signals or we may say a series of binary numbers for digital signals. Digital audio systems represent audio signals in a variety of digital formats. Digital audio data has many digital formats, the most common is .wav file which contains uncompressed audio data. Figure 1 shows the wave in the audio data.

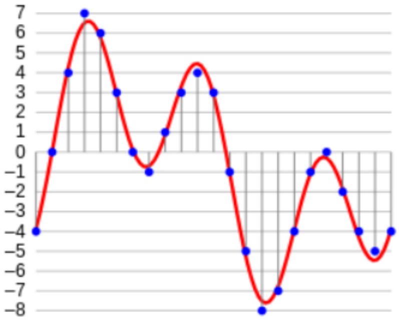


Figure 3.1: Sine wave in the audio file

3.1 Feature Extraction

Feature extraction is a method for identifying essential information features or attributes using different techniques for the same. There a lot of types of feature extraction methods, below given are some of them.

FEATURE EXTRACTION TECHNIQUES:

- Mel Frequency Cepstral Coefficient(MFCC)
- Spectrogram
- Chroma Features
- Spectral Centroid
- Linear Prediction Coefficient(LFC)
- Linear Prediction Cepstral Coefficient(LFCC)
- Line Spectral Frequencies(LSF)

3.2 Feature Extraction Techniques Used

3.2.1 Mel Frequency Cepstral Coefficients (MFCC)

On a contorted frequency scale based on human auditory discernment, MFCC are cepstral coefficients inferred. Windowing the audio signal into partitions divides the discourse signal into outlines is the most important factor in MFCC calculation. Since high frequency formants test lower sufficiency than low recurrence formants, high frequencies are highlighted to ensure that all formants have the same plentifulness. After windowing, the Fast Fourier Transform (FFT) is used to determine each edge's force range. As a result, the channel bank preparation is carried out on the force spectrum using mel-scale. After converting the force spectrum to log space and calculating the MFCC coefficients, the DCT is applied to the discourse signal. The formula for determining the mels for any frequency is,

$$mel(f) = 2595x \log_{10}(1 + f/700) \tag{3.1}$$

In equation 3.1 mel(f) denotes the frequency (mels) and f denotes the frequency (frequency) (Hz).

Equation 3.2 is used to measure the MFCCs:

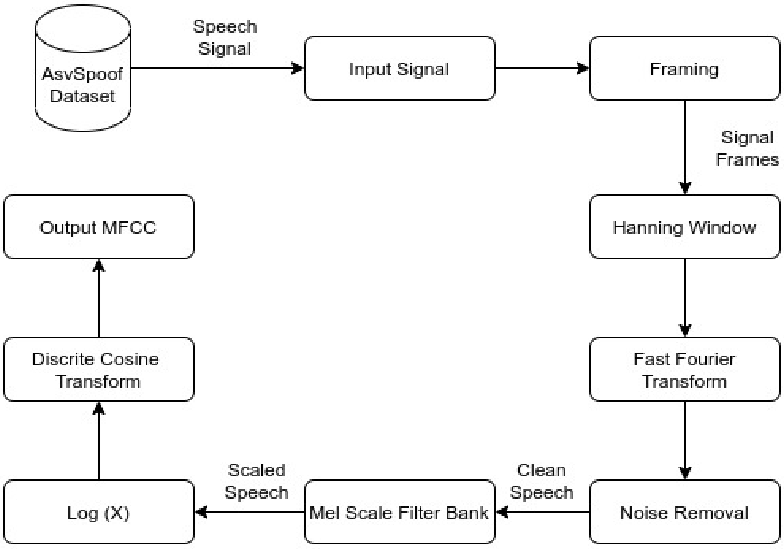


Figure 3.2: Mel Frequency Cepstral Coefficients flow.

$$\hat{C}_n = \sum_{n=1}^{k} \left(\log \hat{S}_k\right) \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{k}\right] \tag{3.2}$$

In equation 3.2 the number of mel cepstrum coefficients is k, the output of filterbank is Sk, and the final mfcc coefficients are Cn.

In some pattern recognition problems involving human speech, cepstral coefficients are said to be correct. They are commonly used in speaker recognition and identification [9]. Below table shows the comparison of different feature extraction techniques.

3.2.2 Spectrogram

A spectrogram is a graph that depicts the frequency spectrum of a recorded audio over time. This means that as the figure gets brighter, the sound becomes more focused around those precise frequencies, and as the figure gets darker, the sound becomes more empty/dead[15]. The fast Fourier transform is a valuable tool for analysing the frequency content of a signal, but what if the frequency content changes over time? This category includes the majority of audio signals, such as music and voice. These signals are known as non-periodic signals. We need a way to see the spectrum of the signal as it changes over time. This is called the short-time Fourier transform, and it is exactly what is done. The spectrogram is obtained by computing the FFT on overlapping windowed portions of the stream[16].

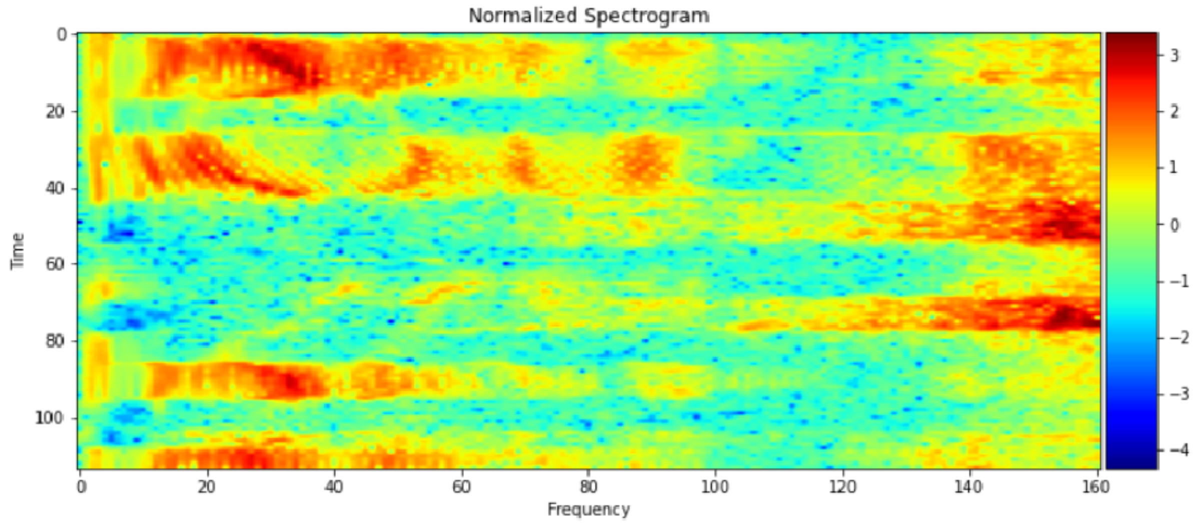


Figure 3.3: Spectrogram

3.2.3 Chroma Feature

Music which has usefully classified pitches (usually into 12 divisions) and tuning that takes into account the scale which is equal-tempered might benefit from chroma-based characteristics, sometimes known as ”pitch class profiles.”. The chroma feature is a condensed descriptor that reflects a musical audio source’s tonal component. As a result, chroma features can be regarded of as a prerequisite for a semantic analysis which is high-level like chord recognition or for example harmonic similarity estimations. A higher quality extracted chroma feature provides for significantly better performance in these high-level jobs. Short Time Fourier Transforms with Constant Q Transforms are utilised to retrieve chroma features.[22].

3.2.4 Spectral Centroid

The spectral centroid is a statistic used in digital signal processing to categorize a spectrum. It depicts the location of the spectrum’s mass center. It has a significant perceptual connection with how bright a sound is perceived. The spectral centroid is frequently related with sound brightness measurement. The ”centre of gravity” is calculated using the frequency and magnitude information from the Fourier transform. The average frequency weighted by amplitudes, divided by the sum of the amplitudes, is the individual centroid of a spectral frame, or:

$$SpectralCentroid = \frac{\sum_{k=1}^{N} kF\left[k\right]}{\sum_{k=1}^{N} F\left[k\right]} \tag{3.3}$$

In equation 3.3 the amplitude corresponding to bin k in the DFT spectrum is F [k] [23].

3.3 Audio Spoofing Attack Detection

Automatic speaker verification (ASV) [13] systems are used for individual verification in a variety of business applications such as call centers, banks, and personal digital assistants (PDAs). These systems, in any case, are vulnerable to spoofing attacks [11]. The vulnerability of ASV frameworks to mocking tackles is a serious problem that needs to be addressed because it poses a real threat to their protection. When used correctly, a mocking assault will give unauthorised access to private and sensitive information. Creating fake discourse, pantomime or mimicry, and replaying discourse accounts are all examples of spoofing attack techniques. To counter such mocking attacks, one may construct a framework that distinguishes between genuine and parody discourse signals; however, what credits could such a framework use to do so? It's safe to assume that observable AI and a proper measure of knowledge would actively seek out such characteristics. In both the ASVspoof 2015 and ASVspoof 20172 public assessment challenges, a few AI frameworks were successful in spoofing assault venue. Frameworks that use deep neural networks are particularly effective at detecting replay attacks (DNNs). Despite the fact that these systems have produced positive results, it is unclear what they have worked out how to do; they are frequently used as a black-box. Is a system that appears to understand a mocking assault really operating with credits relevant to the problem, or is it just a product of how a train/test knowledge base was developed on the other hand.

For example, shows how an edge-based Gaussian Mixture Model (GMM) system prepared for replay caricaturing position on the variant 1.0 of the ASVspoof 2017 data base exploited ancient rarities in the data set to make class decisions. On the refreshed form 2.0 corpus, the same developers recognize a comparable problem for outline-based GMM frameworks. It is still unclear if we'd be able to trust such a system "in nature." Answers to these questions will help strengthen not only the protection of ASV systems, but also the development of new parodying attacks and the improvement of preparing data bases. For all the audio spoofing detection work, the dataset used is the ASVspoof 2017 dataset.

Chapter 4

Siamese Networks

Bromley and LeCun initially proposed Siamese nets in the early 1990s to handle the challenge of signature verification as an image matching problem (Bromley et al., 1993). A Siamese neural network is a type of artificial neural network that employs the same weights to compute equivalent output vectors from two different input vectors. One of the output vectors is frequently precomputed, creating a baseline against which the other output vector is measured. The purpose of the Siamese network is to differentiate between the two inputs X1 and X2. The network's output is a probability between 0 and 1, with a value closer to 0 suggesting dissimilar picture prediction and a value closer to 1 indicating comparable image prediction as output.

4.1 Why is Siamese Network used ?

- In an ensemble, it's nice to have the best classifier: Because it uses a different learning approach than Classification, averaging it with a Classifier can produce significantly better results than averaging two associated

- Supervised models (e.g. GBM and RF classifier) Siamese focuses on learning embeddings (in the deeper layer) that group together related classes and notions as a source of information. As a result, it is possible to learn semantic similarities.

- More resistant to class imbalance: Siamese Networks can recognize a few images per class in the future with just a few photographs for each class thanks to One-shot learning.

4.2 Loss Functions in Siamese Networks

4.2.1 Binary Cross Entropy Loss

The actual class outcome, which can be 0 or 1, is compared to each of the projected probabilities. The score is then computed, with probabilities being penalised based on their divergence from the projected value. That is, how near the result is to the true value. First, let's define binary cross-entropy in formal terms. The negative average of the log of corrected predicted probability is called Binary Cross Entropy.

$$LogLoss = \frac{1}{N}\sum_{i=1}^{N} - \left(y_i * \log\left(p_i\right) + \left(1 - y_i\right) * \log\left(1 - p_i\right)\right) \tag{4.1}$$

In equation 4.1 the chance of class 1 is pi, whereas the likelihood of class 0 is (1-pi).

The first portion of the formula becomes active and the second part vanishes when the observation belongs to class 1, and vice versa when the observation's true class is 0. This is how the Binary cross-entropy is calculated.

4.2.2 Triplet Loss:

A triplet loss compares a base (anchor) intake to a positive (truthy) and negative (falsy) input. The gap between the base (anchor) and positive (truthy) inputs is reduced to the minimum value achievable, the gap between the base (anchor) input and the adverse (false) input is growing.

$$\iota\left(A, P, N\right) = \max\left(\left\|f\left(A\right) - f\left(P\right)\right\|^2 - \left\|f\left(A\right) - f\left(N\right)\right\|^2 + \alpha, 0\right) \tag{4.2}$$

In 4.2 equation, the anchor, positive, and negative images have feature embeddings of fa, fa, fn, with alpha functioning as a margin term to "stretch" the distance variations between like and unlike pairs in the triplet. During the training phase, the model is fed a single image triplet (anchor image, negative image, positive image). The concept is that the distance between both the anchor and the positive images should be less than the distance between the anchor and the negative images.

4.2.3 Contrastive Loss:

Contrastive Loss is a popular and commonly used loss function currently. Rather than the more usual error-prediction losses, it is a distance-based loss. This loss is used for training embeddings that have a small Euclidean distance between similar points and a large Euclidean distance between dissimilar points. The Euclidean distance, Dw, is defined as follows: 4.3

$$(1-Y)\frac{1}{2}\left(D_w\right)^2+(Y)\frac{1}{2}\left\{\max\left(0,m-D_w\right)\right\}^2 \tag{4.3}$$

Our network's output for a single image is Gw, as seen in 4.4

$$\sqrt{\left\{G_w\left(X_1\right)-G_w\left(X_2\right)\right\}^2} \tag{4.4}$$

Chapter 5

Proposed Framework for Audio Spoofing Attack Detection using Siamese Network

5.1 Proposed Approach To Detect Audio Spoofing Attack using Siamese Network

There are many methods giving solutions for detecting the audio spoofing attack instead of audio being a complex data. Various classification algorithms are used to serve the purpose, which proved to efficient too. But using neural network proves to be more efficient in most of the cases. We recently came across a deep neural network that uses two similar networks with same weights to find similarity between two objects or to find difference between the two as it gives comparable output vectors[17], which saves the comparing with all the rest objects. This quality should make it more efficient in the audio domain also. Siamese not being much used in audio domain, remains a problem for getting extremely efficient performance directly so we have different approach for comparing the performance. The work is divided mainly into three parts that have variations:

1. Different Feature Extraction Technique
2. Different Siamese Networks
3. Comparing its performance

Feature extraction techniques used are 1. MFCC (Mel Frequency Cepstral Coefficients): ”Windowing the signal, applying the DFT, taking the log of the magnitude, and then warping the frequencies on a Mel scale, followed by applying the inverse DCT are the basic steps in the MFCC feature extraction technique”[18]. 2. Spectrogram: A spectrogram is a graph that depicts the frequency spectrum of a recorded audio over time [19]. 3. Chroma Feature: Pitch class profiles, also known as chroma-based characteristics, are a strong method for analysing music with meaningfully categorised pitches (typically into dozen categories) tuning that takes into account the equal-tempered scale. and 4. Spectral Centroid: The spectral centroid is a statistic used in digital signal processing to describe a spectrum. It depicts the location of the spectrum’s mass center. It has a significant perceptual connection with how bright a sound is perceived.

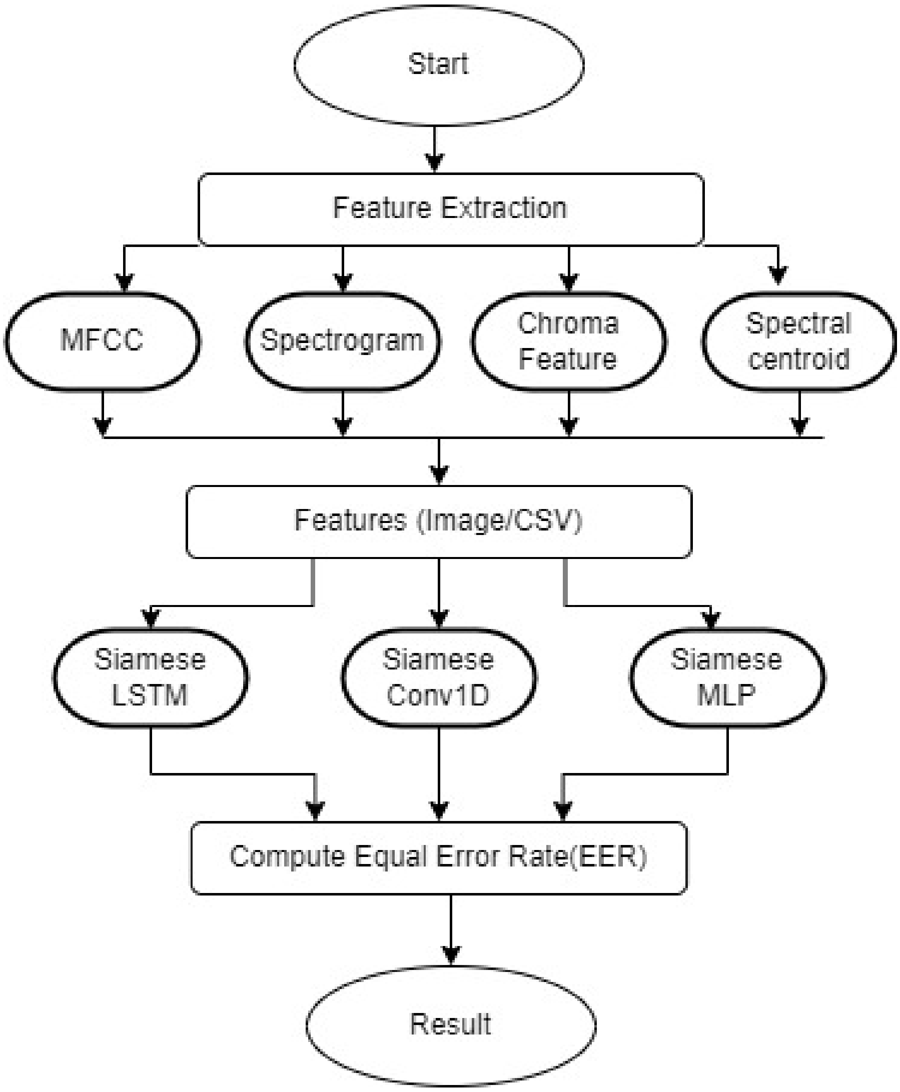


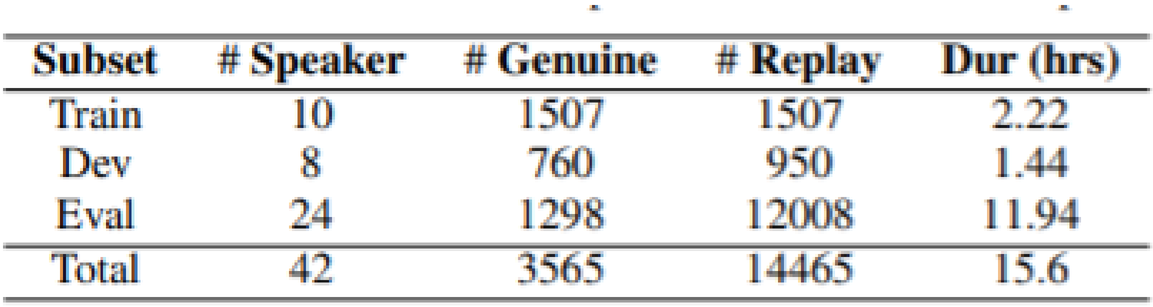
Figure 5.1: Flow for Audio Spoofing Detection using Siamese Networks

5.2 Dataset Description

The dataset that best matches our requirements and is standard is the ASVspoof Dataset. Here we have used the ASVspoof 2017 dataset. The dataset has three directories, training, development and evaluation with different number of audio files in .riff or .wav formats of different duration. All the audio file have 16 kHz of sampling rate and are stored in 16-bit format [20]. Then there is protocol directory which contains three protocol text files in ASCII format for where training consists of list of file that is used to train spoofed and human speech detectors, development with list for trying validation and evaluation with data list to evaluate the final performance of any speech detector. All three files have the same format with seven different columns with the given below labels

- unique file ID
- Genuine or Spoof
- Speaker ID
- RedDots common phrase ID
- Environment ID
- Playback device ID
- Recording device ID

Below Figure has the statistics of the ASVspoof 2017 dataset



Subset	# Speaker	# Genuine	# Replay	Dur (hrs)
Train	10	1507	1507	2.22
Dev	8	760	950	1.44
Eval	24	1298	12008	11.94
Total	42	3565	14465	15.6

Figure 5.2: Statistics Of ASVspoof 2017 dataset

There are ten common phrases used for this, which have there specific ID mentioned in the fourth column, which are spoken by a different speakers. The data labeled spoofed that is, it is recorded and replayed have the data for the fifth, sixth and seventh columns filled unlike the genuine labeled data. The data for spoofed has data for ID for the environment where the audio is recorded, ID of device in which is used to record the audio, and ID of the device used to replay the audio in the 5th, 6th and 7th columns respectively.

5.3 Evaluation Parameters

FAR - The percentage ratio between a valid system invader and the actual number of real intruders using the system. FAR stands for False Match Rate (FMR). FRR - The percentage ratio between the number of actual users of the system and the number of actual users who are rejected or limited for use. False Non-Match Rate is another name for FRR (FNMR) [21]. FAR and FRR can be interconnected by mutually exclusive based on both measures above because FAR and FRR cannot occur at the same time. The link between FAR, FRR, and Equal Error Rate is given in Figure (EER).

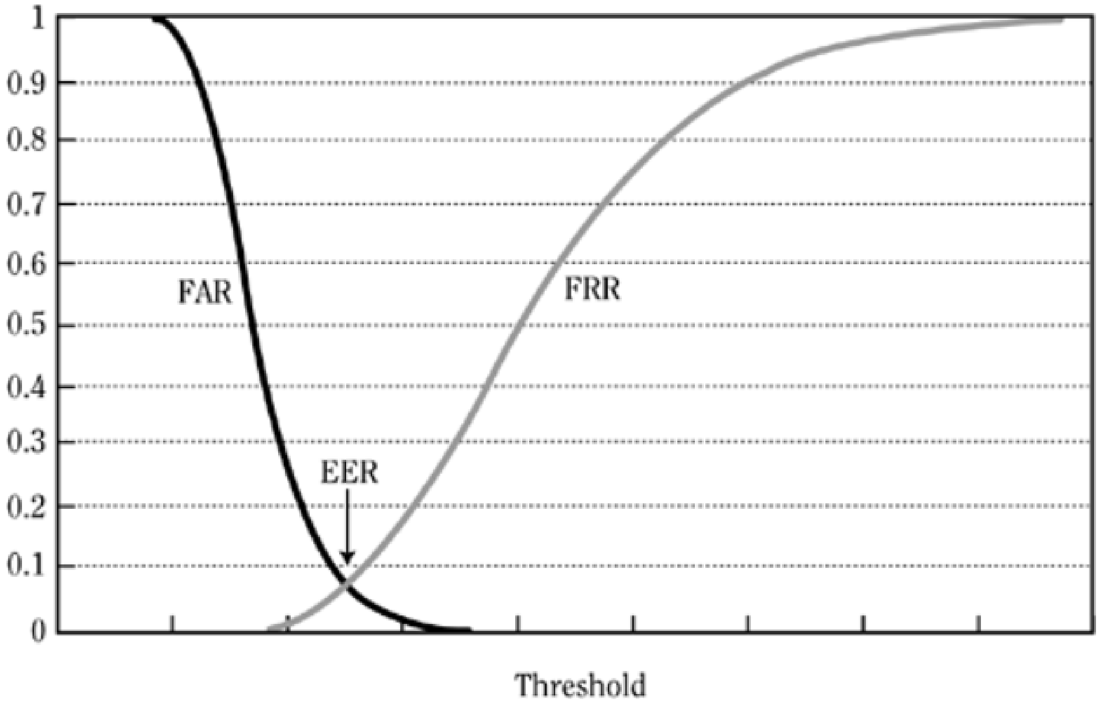


Figure 5.3: FAR, FRR and EER

5.4 Results and Discussion

We tried several different combination of methods with variations in feature extraction methods and classification used with Siamese. The method compares all the combinations with the help of equal error rate(EER) that represents the point which shows the intersection of FAR and FRR where both of them are at the lowest point together that is the threshold value, EER. We tried the below mentioned methods:

- Chroma Feature + Siamese LSTM
- Chroma Feature + Siamese Conv1D
- Chroma Feature + Siamese MLP
- Spectrogram + Siamese LSTM
- Spectrogram + Siamese Conv1D
- Spectrogram + Siamese MLP
- MFCC + Siamese LSTM
- MFCC + Siamese Conv1D
- MFCC + Siamese MLP
- Spectral Centroid + Siamese LSTM
- Spectrl Centroid + Siamese Conv1D
- Spectral Centroid + Siamese MLP

Table 5.1: EERs obtained for different methods used.

	Chroma Feature	Spectrogram	MFCC	Spectral Centroid
LSTM	15	38	4	45
Conv1D	56	50	62	3
MLP	184	13	306	103

The results of the experiments shows that the combination of MFCC + Siamese LSTM and Spectral Centroid + Siamese Conv1D performs much better than all other methods with an EER of 4% and 3% respectively.

Chapter 6

Conclusion

In this paper we have tried using the twin neural networks, that is Siamese Networks with different combinations to evaluate the performance of this very efficient algorithm which is hardly used to deal with complex data like audio. The combinations tried here were evaluated using the performance measure named Equal Error Rate(EER) which is standard method used for measuring performance of audio data. According to the experiments conducted on the ASVspoof 2017 dataset, the performance remains comparable for the rest methods but proves to be very efficient when used the combination of Spectral Centroid + Siamese Conv1D Network with 3% EER and MFCC + Siamese LSTM with 4% EER. Hence using Siamese for audio spoofing attack detection proves to be efficient enough for siamese networks to be taken into consideration to get efficient performance and good results for audio data related researches.

Chapter 7

Future Work

Following are the work that needs to be done in future:

- Other approaches to be tried with yet another feature extraction techniques, other classification algorithms
- Other evaluation parameters will be tried to find more accurate answer by even more thorough performance measures
- Shall generate own dataset with genuine and recorded human voice and try experimenting on that
- Other distances then Euclidean distance can be used to improve the performance

Bibliography

[1] Yuxiang Xu, Guomin Sun, and Jinsong Leng. 2020. Siamese Network for Single Image Rain Removal. In Proceedings of the 2020 4th International Conference on Vision, Image and Signal Processing (ICVISP 2020). Association for Computing Machinery, New York, NY, USA, Article 13, 1–5. DOI:https://doi.org/10.1145/3448823.3448876

[2] Priya Arora and Theodora Chaspari. 2018. Exploring Siamese Neural Network Architectures for Preserving Speaker Identity in Speech Emotion Classification. In Proceedings of the 4th International Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction (MA3HMI'18). Association for Computing Machinery, New York, NY, USA, 15–18. DOI:https://doi.org/10.1145/3279972.3279980

[3] Zachary Langford, Logan Eisenbeiser, and Matthew Vondal. 2019. Robust Signal Classification Using Siamese Networks. In Proceedings of the ACM Workshop on Wireless Security and Machine Learning (WiseML 2019). Association for Computing Machinery, New York, NY, USA, 1–5. DOI:https://doi.org/10.1145/3324921.3328781

[4] L. Vizváry, D. Sopiak, M. Oravec and Z. Bukovčiková, "Image Quality Detection Using The Siamese Convolutional Neural Network," 2019 International Symposium ELMAR, 2019, pp. 109-112, doi: 10.1109/ELMAR.2019.8918678.

[5] C.Wang and G. Tzanetakis, "Singing Style Investigation by Residual Siamese Convolutional Neural Networks," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 116-120, doi: 10.1109/ICASSP.2018.8461660.

[6] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S. Chang and T. Sainath, "Deep Learning for Audio Signal Processing," in IEEE Journal of Selected Topics in Signal Processing, vol. 13, no. 2, pp. 206-219, May 2019, doi: 10.1109/JSTSP.2019.2908700.

[7] T. Kobayashi, A. Kubota and Y. Suzuki, ”Audio Feature Extraction Based on Sub-Band Signal Correlations for Music Genre Classification,” 2018 IEEE International Symposium on Multimedia (ISM), 2018, pp. 180-181, doi: 10.1109/ISM.2018.00-15.

[8] Z. Li and P. Song, ”Audio similarity detection algorithm based on Siamese LSTM network,” 2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP), 2021, pp. 182-186, doi: 10.1109/ICSP51882.2021.9408942.

[9] Abeysinghe, Chamath, A. Welivita and Indika Perera. “Snake Image Classification using Siamese Networks.” Proceedings of the 2019 3rd International Conference on Graphics and Signal Processing (2019): n. pag.

[10] B.Chettri, S. Mishra, B. L. Sturm and E. Benetos, ”Analysing The Predictions Of a CNN-Based Replay Spoofing Detection System,” 2018 IEEE Spoken Language Technology Workshop (SLT), 2018, pp. 92-97, doi: 10.1109/SLT.2018.8639666. doi: 10.1109/ICOT.2015.7498491

[11] S.Shukla, J. Prakash and R. S. Guntur, ”Replay attack detection with raw audio waves and deep learning framework,” 2019 International Conference on Data Science and Engineering (ICDSE), 2019, pp. 66-70, doi: 10.1109/ICDSE47409.2019.8971793.

[12] B. T. Balamurali, K. E. Lin, S. Lui, J. -M. Chen and D. Herremans, ”Toward Robust Audio Spoofing Detection: A Detailed Comparison of Traditional and Learned Features,” in IEEE Access, vol. 7, pp. 84229-84241, 2019, doi: 10.1109/ACCESS.2019.2923806.

[13] Lavrentyeva, Galina Novoselov, Sergey Malykh, Egor Kozlov, Alexander Oleg, Kudashev Shchemelinin, Vadim. (2017). “Audio Replay Attack Detection with Deep Learning Frameworks.” 82-86. 10.21437/Interspeech.2017-360.

[14] (2007). Pitch- and Chroma-Based Audio Features. In: Information Retrieval for Music and Motion. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-74048-3_3

[15] R. Paschotta, article on ’spectrograms’ in the Encyclopedia of Laser Physics and Technology, 1. edition October 2008, Wiley-VCH, ISBN 978-3-527-40828-3

[16] Spectrogram,
https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53

[17] Bromley, Jane; Guyon, Isabelle; LeCun, Yann; Säckinger, Eduard; Shah, Roopak (1994). "Signature verification using a "Siamese" time delay neural network". Advances in Neural Information Processing Systems 6: 737–744

[18] K.S. Rao and Manjunath K.E., Speech Recognition Using Articulatory and Excitation Source Features, SpringerBriefs in Speech Technology, DOI 10.1007/978-3-319-49220-9

[19] Spectrogram
https://towardsdatascience.com/learning-from-audio-spectrograms-37df29dba98c

[20] ASVspoof Dataset,
https://datashare.ed.ac.uk/handle/10283/3055

[21] Yaacob, Mohd Noorulfakhri Syed Idrus, Syed Zulkarnain Wan Ali, Wan Nor Ashiqin Mustafa, Wan Jamlos, Mohd Abd Wahab, Mohd Helmy. (2020). Decision Making Process in Keystroke Dynamics. Journal of Physics: Conference Series. 1529. 022087. 10.1088/1742-6596/1529/2/022087.

[22] Shah, Ayush Kattel, Manasi Nepal, Araju Shrestha, D.. (2019). Chroma Feature Extraction.

[23] https://www.digitalxplore.org/up$_{p}roc/pdf/273 - 14867862539 - 12.pdf$