Tampere University

Muhammad Ahmad Bashir

VOICE SPOOFING COUNTERMEASURE

Securing Automatic Speaker Verification Against Logical Access Attacks

Faculty of Information Technology and Communication Sciences

Master of Science Thesis

June 2023

ABSTRACT

Muhammad Ahmad Bashir: Voice spoofing countermeasure securing automatic speaker verification against logical access attacks.

Master of Science Thesis

Tampere University

Master's Degree Programme in Data Science

June 2023

The automatic Speaker Verification (ASV) system performs verification of customers based on the user's voice. Mostly, ASV systems are susceptible to voice-based attacks, namely, logical and physical access (LA and PA). The detection of the above attacks is now significantly improved. Whereas attackers can utilize augmented devices to record the consumer's speech and replay it besides ASV to gain access for unethical purposes. The major objective of this approach is to create a powerful voice anti-spoofing measure for identifying voice-spoofing attacks. Specifically, the security of the ASV is enhanced by proposing hybrid feature set containing salient information and implementing countermeasures against voice spoofing attacks. In this work, the voice signals are represented using four features, specifically Mel-frequency cepstral coefficients, gamma tone cepstral coefficients, spectral skewness, and spectral energy. The classification is performed using deep learning classifiers such as long-term short-memory networks. In this work, voices features are examined several and their ability to obtain crucial information from voice signals and discriminate between authentic and fake voice samples. Furthermore, the superiority of the system over traditional classifiers using various machine-learning algorithms is demonstrated. Specifically, the system achieves 100% accuracy, 99.9% precision, and recall and F1 scores of 100% and 99%, respectively. In addition, the EER is 0.01%. Consequently, the public dataset ASVspoof2019 LA is used for experimentation.

Keywords: Automatic Speaker Verification, Forensics, ASVspoof2019, Logical Access.

PREFACE

This is to certify about my research work titled "*Securing Automatic Speaker Verification Against Logical Access Attacks"* is an original and my proposed work. I have never presented this research anywhere else, and the content/materials used have been referred to.

I'm grateful to my supervisor, Dr. Martti Juhola, Professor, of Data Science, in Finland, for his precious supervision and support during the research progression. His knowledge and supervision have been influential in the successful accomplishment of this thesis.

I am also grateful to my friends and family for their steadfast backing, thoughtfulness, and inspiration and I also extend my gratefulness to all who have facilitated me in this drive, comprising my colleagues, instructors, and classmates.

This thesis is dedicated to all of them, and I hope it will make a meaningful contribution to the field of Data Science.

Tampere, 7th June 2023
Muhammad Ahmad Bashir

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ASV	Automatic Speaker Verification
VC	Voice Conversion
EER	Equal Error Rate
KNN	K nearest Neighbors
MFCC	Mel frequency Cepstral Coefficients
GTCC	Gammatone Cepstral Coefficients
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
TEO	Teager Energy Coefficients
GMM	Gaussian Mixture Model
DL	Deep Learning
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
CQCC	Constant Q-Cepstrum Coefficient
ETEO	Enhanced Tiger Energy Operator
SVM	Support Vector Machine
LSTM	Long Short-Term Memory Network
RAD	Rapid Application Development
TLC-AM	Transmission Line Cochlear - Amplitude Modulation
TLC-FM	Transmission Line Cochlear - Frequency Modulation
ERB	Equal Rectangular Bandwidth

1. Introduction

1.1. Background of the Study

Communication is a necessary part of life, and several ways of communication have been introduced nowadays. But communication through voice has always been proven as the most effective way of communication.

Two basic things are required for voice communication i.e., listening and speaking. Voice communication is considered the simplest form of communication, as there is no need for any electronic gadget for this type of communication. Every person has a different voice from others, so it can be used for authentication purposes such as biometric identification. Voice recognition can be used as a simple measure for biometric authentication. Moreover, this type of authentication does not require any advanced equipment such as modern sensors. It is possible to use a simple microphone for this purpose. The simple term used for voice biometrics is “speaker recognition process”. Generally, this process is recognized as a one-to-one process. In this process, the voice of one person is compared with the voice of another voice. If both voices are the same, then it is called speaker verification. But, this is a one-to-many process, in which one voice is compared to many other voices. However, this process eventually leads to several recurrences of one-to-one comparisons.

Many companies are using voice biometrics for authentications. For example, ING Bank (located in the Netherlands) has introduced a voice-recognition mobile application for customer authentication. According to the bank report, more than hundred thousand people use voice biometrics in this bank for their authentications and other online bank transactions. This company thinks that most companies will shift to voice biometrics soon because it is a simpler and easier way of authentication rather than other previously used methods. The director of the internet of ING bank has stated that their mobile application has other methods of authentication, but most of their customers prefer to use voice recognition for the authentication. Moreover, he said that the company is trying to provide more user-friendly voice features in their mobile app for their customers. This company provides various facilities through mobile apps such as account login, check balance, online payments, funds transfer, and transaction history. These aforementioned operations can easily be performed by the customer’s voice commands within a short time frame. The customers can interact with the mobile app through their voices, and it is the same as the

customers interacting with other people in the bank branch to perform operations on their bank accounts. For example, the customer can give voice commands to the mobile app for giving the location of the nearest possible bank branch or to initiate any online transaction. This app has already enabled the feature of initiating any online transaction and login authentication of customers through their voices. The updated version of this app has both features. The login authentication is performed based on voice biometrics. At the start, the system asks to register the customer voice to initiate voice authentication. This registered speech is scrutinized for several distinctive forms that are matched to the voice identity verification on the file.

Conferring to the bank administrators, this voice feature app was first introduced by ING bank in Europe. The key intent of this app is to provide the most effective way of online banking, which should be similar to physical banking. Another official of the bank stated that customers do not like to rise daily and go to the banks for their transactions. Eventually, the voice feature app will replace the other transactions authentications such as PINs and passcodes, etc. Because the voice feature app is easy to use for authentication and other transactions rather than questions asked by the banks to ensure customer security. In the backend of this app, there are several characteristics of voice that are determined for initiating the authentication process of any speaker. These voice characteristics are comprised of pitch, loudness, intensity, and harmonic structure. The company also stated that there are a lot of their customers that are using the voice feature of the app. Customer feedback identified that the voice feature of the app takes less time and is also easy to use. This method of authentication is brand new in mobiles as this is not used by any mobile app for authentication purposes.

Another bank named Barclays has introduced several advanced methods of authentication such as voice recognition, fingerprint, facial scanner, etc. However, it has been seen that other authentication methods are not getting popular, instead the use of these methods has decreased after the development of voice feature authentication. More precisely, other authentication methods do not real the actual issue or cannot be installed in a real-time environment. This detailed discussion about online banking with the voice feature has proven very useful, but the security of online banking is still a question as hackers have started using advanced equipment to breach security, which has compromised the security of the customer’s data.

Nowadays, the world has become advanced in every field of life. Modern equipment and gadgets are introduced daily. This advancement has increased security risks, therefore, there is a need for more secure and unbreachable authentications of the customers to avoid any type of security threats. Initially, cryptographic methods were used for authentications in the form of passwords or passcodes. Many times, if the access became suspicious then the customers had to give both passwords and cards to prove their identity for accessing their accounts. The main disadvantage of this system is that the customers must remember their passwords all the time when they want to log in to their accounts to perform transactions, checking history or unlocking their phones, etc. Advanced methods have introduced the modern features of fingerprint, facial recognition, and voice recognition. These methods do not need to remember any type of password. These above-mentioned features are named biometrics. Although these modern methods may still have security risks, these methods have the advantage of not remembering their passwords. Furthermore, biometrics are accessible all the time and cannot be stolen. Therefore, these biometrics are an automatic and digital identification system in which the characteristics of an input signal are matched with the already signal saved in the database. There are different modes by which biometric systems can be utilized. The following section discussed these modes of biometric systems in detail.

1.2. Problem Statement

When the ASV verifies the customer, it approves or rejects the speaker's identity claim. This type of authentication uses the audio of human beings, and this way of communication is a very common authentication method for allowing users access to the system. Different types of spoofing attacks are possible such as PA, and synthetic or LA attacks. The physical access attacks are related to playing audio tracks beside the ASV systems and logical attacks are related to generating audio similar to the voice of the original customer with the help of advanced techniques and algorithms. Modern devices have low error rates, and hackers use these modern devices or advanced spoofing algorithms to produce a high-quality voice that appears to be very similar to the original voice of the customer. It has become a necessity to identify the original and spoofed voices. These circumstances are creating significant risks to the security of the ASV systems. The most common way of identifying users is their voices in day-to-day life. Voice is used for customer authentication and can improve ASV security with automated identification and spoofing voice

detection. Research studies have worked to detect voice spoofing attacks. Even so, the issue is still unresolved, and it is essential to propose powerful anti-spoofing techniques that can identify fake audio to protect ASV systems. In this research study, the primary purpose is to build a voice spoofing countermeasure to detect logical access attacks.

1.3. Aims and Objectives and Research Questions

This section has details of the aims and objectives as well as the research questions.

1.3.1. Aims and Objectives

- Aim is to develop new robust voice anti-spoofing measures to recognize fake speech based on acoustical algorithms.
- The major goal of this approach is an examination of acoustical algorithms/features for developing a system based on hybrid features to detect fake voices.
- The goal of this study is to examine traditional ML classifiers for spoof identification performance.
- The objective of this study is to examine the deep learning (DL) classifiers for their performance in logical access attacks.

1.3.2. Questions and Hypothesis

- Do the hybrid features comprise the greatest amount of details to identify a powerful attack type, namely, LA spoofing?
- Do the conventional classifiers work efficiently for powerful LA identification?
- Does the DL-established technique confirm higher functioning than the conventional classifier?

1.4. Contributions

The major contributions of this study are given below,

- Proposed a novel hybrid features that obtain maximum details from the audio samples.
- Conducted extensive experiments on the publicly available dataset.
- The proposed spoofing countermeasure successfully detected voice spoofing attacks using a deep learning classifier, LSTM.

1.5. Signification of Work

This study will propose powerful countermeasures for the security of ASV systems. Countermeasures protect ASV techniques from hackers gaining suspicious and illegal access. The hackers will not be able to breach the security of the novel ASV system because it will examine the audio features and pass them to the DL model for categorization purposes. The ASVspoof organizer baseline used constant Q cepstrum coefficients and GMM for the classification. Whereas their developed system is unreliable to utilize in a real-time environment because its capabilities do not capture the maximum amount of information. Hence, it inspired us to utilize hybrid features to achieve this goal.

1.6. Scope and Limits of this Work

The main scope of this study is to identify the audio spoofing attacks and increase the security of the ASV system for login and other authentications. Voice authentication is a biometric technique that identifies customers by measuring the difference between innumerable voices. For security, since passwords can be forgotten or insecure, speech recognition systems allow users to use their voice as a password. Speech recognition technology works by digitizing human voices and generating templates named “voice print”. Speech recognition technology divides voiceprints into a number of segments composed of formants. A formant consists of a number of tones that identify the user’s voiceprints. Voiceprints are stored in the database just the face and fingerprint scans.

There have been several disclosed incidents of hackers exploiting vulnerabilities in voice recognition systems to break into homes and businesses to harm their owners. Modern ASV systems can be breached, and there is a possibility of getting access to the systems by using spoofing audio. This study focuses on examining the audio signal of the acoustic features that capture the maximum characteristics of the voice signal. In addition, the scope of this work is to adopt a DL model that can efficiently identify the spoofed and original voices.

1.7. Tools

MATLAB 2022a is the tool that will be used for the implementation of this research study. Matlab2022a has several audio editing tools. It is easily possible to perform feature extraction and classification with the MATLAB tool.

1.8. Workflow

The details about the working of the proposed system are described in this section. The proposed system aims to classify original and fake audio. Fig 1.1 shows the workflow of the system. Firstly, the dataset of ASVspoof2019 is used. After that, features are extracted, followed by acoustic feature classification. In the end, the system identified the voice signal as authentic or fake. Fig 1.1 depicts the workflow of propsed system.

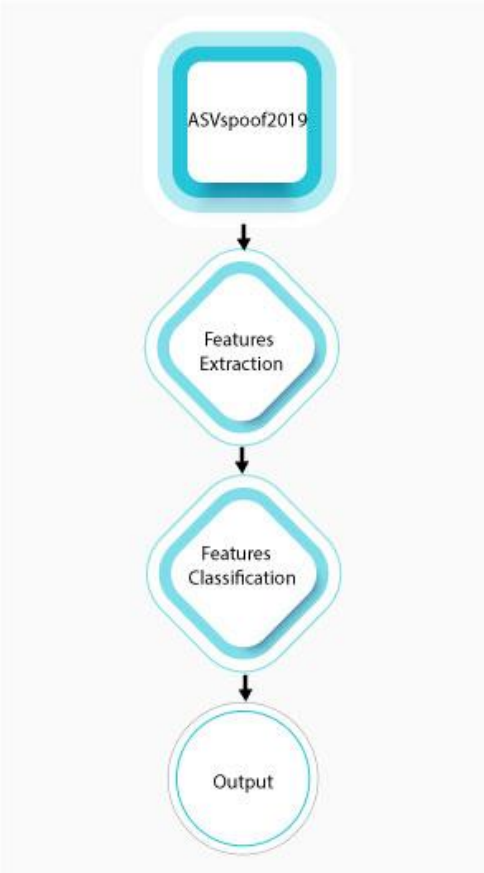

Fig 1.1: Proposed workflow.

1.9. Research Method

The main objective of this research study is to identify voice spoofing attacks against ASV systems. The implementation consists of two basic phases: the feature extraction phase and the classification phase. In the first stage, two features of 14, 14-dimensional, and the remaining two features of 1, 1-dimensional are extracted, named MFCCs and GTCCs, Spectral Skewness, and Spectral energy.

After that, DL classifier is used, i.e., LSTM. In the end, the proposed anti-spoofing measures determined the validity of the client based on the provided audio signal. All experimentations employed the ASVspoof2019 dataset. Information on the proposed working structure is given in Chapter 3 and Section 3.1.

1.10. Feature Extraction

This segment provides an analysis of the features utilized to propose anti-spoofing measures. The proposed system uses four important voice features, MFCC, GTCC, spectral skewness, and spectral energy for speech representation. The information on the features obtained is given in these sections.

1.11. Structure of Thesis

The remaining thesis is structured as the second chapter gives detail about relevant studies, the third chapter has a detailed discussion of this approach, the fourth chapter has details of investigational evaluation and finally, the last chapter has the conclusion of this work.

2. Literature Review

This chapter has discussed the current advanced methods to detect voice spoofing attacks.

2.1. Modes of Biometric Systems

There are three modes of biometric system, namely, verification, identification, and screening. All three modes are discussed in detail in subsequent sections.

2.1.1. Verification

The verification is related to the validation of the customers. In this phase, the system performs the one-to-one processing by comparing the input signal with the signal saved in the database for a specific and giving access to the system to the respective customer. The verification is used for multiple purposes such as login credentials, giving access to several actions, and accessing mobile phones or other systems.

2.1.2. Identification

The identification process performs the one-to-many comparison by comparing the input signal with all signals saved in the database. The identification process identifies exclusive of several individuals retrieve of distinctiveness from the employers of the organizations. The recognition is used for restricting access to a person from multiple accounts or identities. The identification process in identity cards, driving licenses, and other licenses.

2.1.3. Screening

Screening is also a type of identification in which a one-to-many comparison is performed by the recognition system to identify whether a person belongs to the watch list or not in the database. The screening is performed in many places such as airports, surveillance, and other security places.

This research study aims to develop an automated system for the speaker verification of the customer to avoid voice spoofing attacks. The one-to-one comparison technique is used for validating the input voice signal with the signal already saved in the database of the system. This type of system is known as automatic speaker verification (AVS) system. The ASV systems can face two types of cyber-attacks, i.e., zero-effort attacks and spoofing attacks. The details of both of these attacks are disused in the below section.

2.2. Types of Attacks

This subdivision discusses the categories of attacks in detail.

2.2.1. Zero Effort Attack

This is a simple type of ASV attack in which an unregistered person speaks in the voice of a real customer to get access to the customer account. This type of voice is always malfunctioning and is easy to detect. The system performs the one-to-one comparison and detects this type of attack as the system failed to identify the user.

2.2.2. Spoofing Attack

In a spoofing attack, fraudsters employ fake speech of users against authentication systems. This attack makes the authentication system more vulnerable because advanced technologies can store the speech of authentic users. Spoofing attacks are further divided into 4 classes, namely, impersonation, voice play-again (replay), speech synthesis, as well as converted speech attack. Each of these attacks is discussed as follows.

2.2.3. Impersonation

It is the simplest type of attack in which a person tries to mimic the voice of the original customer. This attack is straightforward to identify as the properties of the mimic and original voice are easily distinguishable.

2.2.4. Replay

The type of attack in which the recorded vocal sound of the original customer is performed before ASV systems are called a replay attack. It is a difficult type of attack as it becomes difficult to discriminate between the original voice and the recorded voice. Moreover, the system must detect the various characteristics of the recording equipment such as the microphone and recorder.

2.2.5. Text To Speech

There are various techniques developed for converting written text to the speech of the customer by using his other audio. This is a modern type of attack, and it becomes difficult for the system to detect text-to-speech attacks.

2.2.6. Voice Conversion

An attack in which the recorded voice of an unregistered person is altered in such a way that it sounds like the voice of the original customer without changing the context of the message is called a voice conversion attack.

Furthermore, voice spoofing can be described as both physical access and logical access attacks. The physical access attack uses some kind of equipment such as a microphone, whereas the logical access attack uses algorithms for vice spoofing. In the last few years, studies have focused on developing anti-spoofing systems. The studies showed that the already developed techniques can be improved for increasing anti-spoofing efficiency.

These spoofing attacks are affecting the efficiency of ASV systems. The researchers have started developing modern anti-spoofing ASV systems to remove or decrease the risk of these attack threats. Voice spoofing has become a popular topic for researchers, and the research community has focused on the ASV systems as observed by ASVspoof 2015. Logical access attack is the main issue discussed these days in which modern techniques and algorithms are used for voice conversions attacks and text-to-speech attacks. The challenges for both PA and LA are discussed by BTAS 2016. ASVspoof2017 described challenges related to physical attacks, as these voices were recorded under environmental circumstances such as noise. ASVspoof2019 also described challenges related to both physical access attacks and logical access attacks. In this challenge, several thousand audio attacks were generated by using modern techniques and algorithms. Furthermore, ASVspoof2021 challenge has been introduced that contains both physical and logical access attacks. Moreover, ASVspoof2021 has also developed a modern deep fake along with the two previous fake attacks. The deep fake attacks use the audio data of the targeted people from different platforms such as social media.

Spoofing identification is a binary classification task that has the objective to differentiate between original and fake voices, e.g., there are two predictions for each audio such as whether the audio is original or fake. The modern machine learning algorithms used for voice spoofing detection include the Gaussian Mixture Model (GMM) and Neural Network (NN). The scientific society has developed several audio features such as spectrogram, linear frequency cepstral coefficients, constant Q cepstral coefficients, and ambiguous audio samples to train GMMs and NNs. The previously mentioned challenges also used the NNs-based methods for spoofing detection and

achieved significant results. There are specific agreements used for checking the efficiency of presentation attack detection systems. ASVspoof2019 makes use of the usual standards together with identical mistakes charge (EER) and min-tDCF are utilized to check the comprehensive presentation of the countermeasures ASV systems. The basic standard employed in the ASVspoof2021 is EER, which is utilized for both PA and LA attacks in addition to voice deep fake venture.

Anti-spoofing measures to enhance the security of ASV systems should be hardened and advanced not only to distinguish the classes of attacks defined while the model's training, but also to detect invisible attacks with lesser EER rates. These challenges were addressed by, in which NNs were used for extracting audio features to enhance the recorded voices. The developed anti-spoofing systems are used to determine the authentication of the voice signals. The next phase after extracting the feature is to classify the data of original and duplicate audios. Research communities have utilized either ML classifiers or deep learning (DL) classifiers, and DL classifiers have manifested amazing results. The features extracted by DL can be classified into two categories such as utterance-level features and frame-level features. Moreover, the types of features of the speech signal input provided to the DL feature extractor can check the effectiveness of anti-spoofing. Therefore, the studies have developed different features for identifying the spoofing of the voices such as magnitude-based voice feature, phased-based voice features, and features used with raw samples of voices. Significant results have been achieved by using DL features at frame level and spoof identification, e.g., the x-vector is more famous in ASV systems due to its superior performance as compared to the i-vector. For spoofing protection, Deep Neural Networks (DNNs) and Convolutional Neural Networks (CNN) are used to acquire deep features at the frame level, and the layers of the CNN have the powerful ability to identify features caused by spoofing algorithms that are utilized in logical spoofing attacks. In a noisy environment, CNNs can be viewed as a type of filter bank, and their filters are improved for the specific task of identifying spoofs. Moreover, CNN based residual neural network was utilized in identifying the spoofing attacks.

There is a need to fit the features of the frame level in an individuality vector that describes the whole representation. There are various methods applicable for merging the characteristics such as averaging, statistical pooling, and Recurrent Neural Networks (RNN). For spoofed audios,

RNNs have proven to be most efficient at obtaining features and retaining temporary artefacts. The study in [1] used a Gated Recurrent Unit (GRU) to combine CNNs with RNNs and evaluate logical access attacks by bringing out informative deep features. Another study in [2] used a combination of multiple fully connected layers with two Long Short-Term Memory (LSTM) for developing an entire system. Likewise detected physical access attacks by combining RNNs with light CNNs. Additionally, another study developed an improved DNN based on CNN for identifying physical access attacks [3]. Recently, there are various RNN-focused end-to-end systems introduced in [4], and [5] to identify physical access attacks.

The main aim of anti-spoofing research has been on the system's functionality in isolated environments, and only a few studies have been conducted that considered the affecting environments such as audio noise. Real-time systems have the biggest issue of noise that decreases the efficiency of the system. There are several types of effects that noise has on different types of spoofing attacks. Hence, the type of attack can be identified by the noise of electronic gadget playing in the background. This is a complicated process to identify this type of noise. For example, the TTS anti-spoofing system developed in [6] behaves differently in harmonic and noisy environments. Its efficiency degrades rapidly in noisy environments. The main idea of this research work is to overcome this lack of robustness.

Initially, [7] studied the influences of noise on anti-spoofing systems and evaluated the performance of original characteristics in earsplitting surroundings. Additional NNs established countermeasure in [5] was assessed in 5 various noises with echo circumstances. In [5], the performance of countermeasures in noisy situations has been significantly improved. This upgradation demonstrates that the discriminating function of NNs can be learned even with noisy data. In addition, [5] Noise Aware Training (NAT) helped to improve robustness in noisy circumstances.

2.2. Related Work on the Voice Spoofing Attacks

A detection method was proposed in [7] for speech playback spoofing using glottal excitation and high-frequency band. The researcher determined the importance of glottal detail and size-based features of speech signals for replay attack identification in an ASV system. The details on glottal excitation were obtained using an iterative adaptive inverse filtering procedure showing different

details of original and duplicate voices. Glottal information was extracted by applying the above method to discard the effects of the acoustic tract and lip shine using both inverse filtering and integration. The researcher used constant-Q cepstrum coefficients (CQCC) to obtain features from glottal excitation spectra. A decrease of 3.68% and 8.32% with the same error rate was noticed for development and evaluation sets.

In [8], various features such as enhanced Tiger Energy cepstral ratio, signal mass, and Enhanced Tiger Energy Operator (ETEO) have been studied for identifying duplication attacks. Audio signals have been identified to have sound reproduction properties in all regions of the spectrum. The researcher examined ETEO features relevance using several approaches, including amplitude and frequency modulation, mathematical evaluation, and detailed analysis of spectroscopic examination. Moreover, the comparison between the already extracted features and EETO was performed to grade features. The classification of the spoofing and non-spoofing audios was carried out by using three main classifiers named Light-CNN, CNN, and Gaussian Mixture Model (GMM).

The study in [9] studied a spectrum analyzer called Cochlea because it contained both sharp frequency adjustments and compressions. The researcher developed a technique by utilizing adaptive notch and resonant filters in this model. The analyzed non-symmetrical filter was redesigned with an increase in frequency and level-dependent features.

In [10], the researcher proposed a secure ASV model for detecting played audio, to protect the ASV system from unauthorized access. The classification of spoofed and non-spoofed audio was performed using an ensemble classifier. Furthermore, the model was examined based on two datasets: ASVspoof2019 and a voice spoofing identification compilation developed specifically for replay attacks. This model also helped in identifying the logical access attacks that are caused by the spoofing algorithms.

In [11], a model for the identification of voice spoofing attacks was developed that was based on various classifiers such as SVM and GMM. Additionally, a feature was developed for the Mel sub-band energies based on linear prediction residual evaluation. The features were extracted using a novel feature extraction method that was based on linear prediction coefficients. The researcher used ASVspoof2017 for the experimental setup, which produced an EER of 4.8% on the assessment data.

In [12], the researcher introduced the LiveEar model that identified spoofing voices to protect voice assistants such as Cortona, Siri, etc. This model was developed based on the difference in distance position between playing voices and the live voice of any human being. Especially, the method used the differentiations in arrival times in samples of phonemic voices. The classification was performed using SVM for differentiating between original and duplicate voices.

In [13], the researcher proposed an audio duplication identification method based on LTP and GTCC features to identify spoofed voices and protect cyber-physical systems and IoT devices. The above-mentioned features were combined and given to the SVM classifier for classification. This article also examined multiple duplication attacks. The experimental setup contained two types of datasets, i.e., ASVspoof2019 and VSDC. VSDC dataset contained multiple sequences of duplicate attacks, but the older dataset contained only voice samples of normally played voices.

In [14], the author proposed a voice duplication identification model using two types of features: spectral and spatial signals. The study mainly focused on non-speech parts instead of speech parts where people speak, but electronic gadgets played audio noise or other electromagnetic noises. Moreover, this study utilized spatial features that were based on generalized cross-correlations to observe the differences. To improve the efficiency of the system, the resulting values of both spatial and spectral properties were combined, especially the generic cross-correlation and spectral properties.

In [15], the identification of spoofing attacks was enhanced by using two dissimilar features such as logSpec and cepstral coefficients. A linear prediction signal was used to obtain these above-mentioned features. It has been noticed that these features demonstrated collaborative dissimilarities of voice regions and acoustic glottal flow. Two spaces were analyzed for their importance, i.e., t-stochastic and modelling space. EER reductions of 7% and 51.7% were calculated for the development and evaluation sets.

In [16], voice duplications were identified by using an algorithm of energy separation. The researcher used MFCC and instantaneous amplitude as well as many more features in this model. Features obtained from the voice dataset were input to the GMM to differentiate between spoofed and non-spoofed voices. To avoid the noise problem, this study also investigated various teager energy operators. The experiment was performed in isolated as well as external environments to

verify the effectiveness of the utilized technique. In terms of EER values, enhancements of 21.88% and 66.34% were noticed in isolated and noisy environments.

In [17], a live identification technique was developed for voice inputs based on the constant-Q transform with geometrically distributed frequency bins. This study used pop sounds for checking whether the person speaking in the microphone is present at that moment. The study identified the properties of sound at a very low frequency.

In [18], voice spoofing was identified by using the combined glottal-MFCC and shifted-CQCC. The MFCC was applied to glottal input for obtaining GMFCC features, and CQT was applied for obtaining the shifted-CQCC signals. Shifted-CQCC produced an 11.34% EER and GMFCC produced a 7.94% EER value. Furthermore, the combination of CQCC and GMFCC produced an EER of 8.60%.

In [19], an anti-spoofing method was developed for smart homes called ARRAYID. It is used to collect voice data instead of using sensors for differentiating between recorded voices and live human voices. In this article, the researcher also investigated liveness detection array fingerprinting using a recording gadget called a microphone array.

In [20], various neural networks were used for the identification of spoofing such as CNN and LSTM. The study also explored spectral features named CQCC. Additionally, the framework used both dynamic and static CQCC for rendered audio. A time-distributed wrapper was utilized with LSTM as the backend classifier. This study used two layers for spoof identification. In the first layer, neural networks were utilized, whereas, in the second layer, LSTM and time-distributed wrapper were. The results showed that the neural networks proved more significant for the identification of spoofing attacks.

In [21], a framework was developed for differentiating between the spoofed and non-spoofed voices. The focus of this study is to find relevant features that are caused by recording devices such as microphones. In addition, this research used harmonics and DL-RAD functions, especially high- and low-frequency harmonics, along with harmonic energy ratio, low spectral ratio, low spectral dispersion, and low spectral difference dispersion to achieve a variety of original and duplicate voices. The classification was performed using an SVM classifier.

In [22], the main aim was to identify the spoofed voices by analyzing the microphone and speakers. A familiar voice spoofing procedure was also analyzed for proposing this anti-spoofing system. In

this system, each recording device has an attribute of ASV, the original voice was passed to the recording device once, but there was the possibility of an attack, so the duplicate voice produced for the planned attack was passed twice to the same device. The previously developed identification systems were unable to identify this type of spoofing attack.

In [23], another anti-spoofing system was developed based on cepstral features. The researchers carried out sub-band analysis for those cepstral features. ASVspoof2017 was utilized for experimental purposes. EER value of 36.33% was improved for this system.

In [24], an anti-spoofing measure technique was developed for duplicate voice inputs based on the cochlear model. The basilar membrane is shown as an avalanche of filters with decreasing resonant frequencies. Furthermore, the researchers developed two features, i.e., Transmission Line Cochlear - Amplitude Modulation (TLC-AM) and Frequency Modulation-based (TLC-FM). Both features were used to obtain features from the adjustments. TLC-AM resembled the generation of inner hair cells and could exploit the AM properties of voice signals. In addition, TLC-FM was extracted by deriving the in-phase and out-phase voice signals. The results show the combination of TLC-FM and TLC-AM features, which surpass the filter bank. Radiofrequency selection proved useful for identifying duplicate voices.

In [25], the study demonstrated the value of recognizing duplicate voices by using linear prediction signals. Playback devices have non-regular frequency responses driving the input acoustic signals, causing spectral defects of the recorded voice signal. The study found that mostly the properties of voices get affected that are below 300 Hz. These properties are contained by the linear prediction analysis. Furthermore, the other two features were extracted and combined for identifying the duplicate voices. These features were used as input to the GMM model for classifying the original and duplicate voices. The comparison showed that the results of residual MFCC surpass the baseline system.

In [26], another spoof identification system was developed to identify spoofed and non-spoofed voices. This system utilized the linear frequency residual cepstral coefficients. The researchers studied the properties of harmonic excitation sources that are accessible in linear prediction signals. The classification was performed using CNN and GMM classifiers for differentiating between the original and duplicate voices. The researcher observed a decrease of 28.777% and 42.72% in EER values for both datasets.

Moreover, the researchers carried out feature score-level integration by utilizing a specific classifier. ERR value of 2.40% was decreased on the training dataset, whereas ERR value of 9.06% was decreased on the testing dataset.

3. Proposed Methodology

3.1. Approach

The main goal of this study is to propose a system for the identification of voice spoofing attacks on ASV systems. The implementation consists of two basic phases: the feature extraction phase and the classification phase. In the first stage, two features of 14, 14-dimensional, and the remaining two features of 1, 1-dimensional are extracted, named MFCCs and GTCCs, Spectral Skewness, and Spectral energy. After that, a deep learning classifier, i.e., LSTM is used. In the end, the proposed anti-spoofing measure determined the validity of the client based on the provided voice input signal. All experiments utilized the ASVspoof2019 dataset. Fig 3.1 demonstrates the anti-spoofing measure developed for the ASV systems.

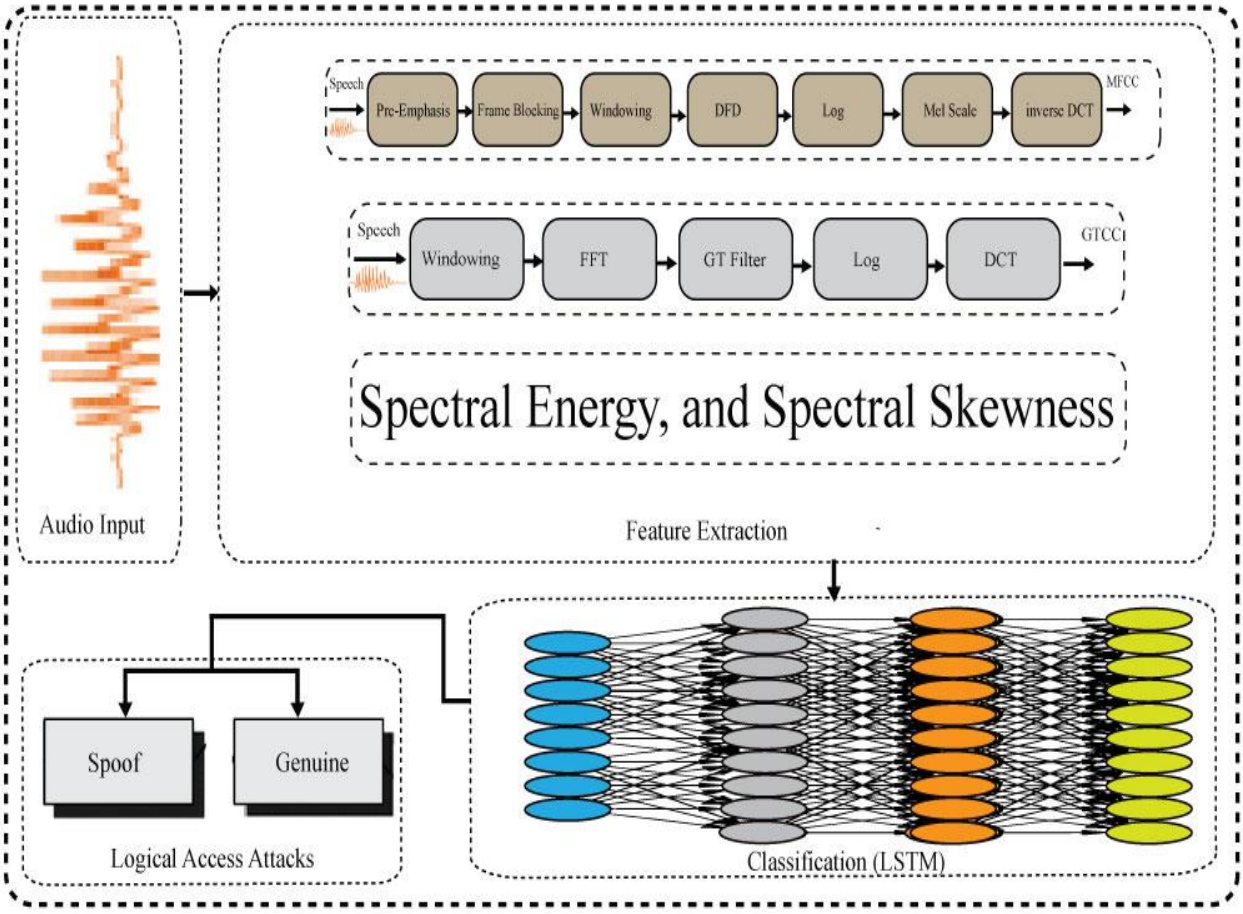

Fig 3.1: Proposed working mechanism.

3.2. Implementation

MATLAB 2022a is the tool that is used for the implementation of this research study. Matlab2022a has several audio editing tools. It is quite easy to perform feature extraction and classification with the MATLAB tool.

3.3. Data in and Out

This section explained that what are the inputs required by the proposed system, and what the output system will provide to the user. In this study, the experiments were carried out on the basis of a publicly available dataset named ASVspoof2019 LA. This dataset contains the voices of 40 people. So, the system used these voices as input and classified the given voice in authenticate and spoof category as output.

3.4. Mel Frequency Cepstrum Coefficients

Feature extraction from MFCC is a famous technique for extracting features from voice signals. This is also called a filter bank-based technique for extracting features in the cepstral domain. In the early stage of MFCC feature extraction process, pre-emphasis is done for an increasing amplitude of high frequencies followed by frame blocking, in which signals are further divided into block for better analysis. Next, windowing the signals is applied for keeping the continuity of the signals, which is followed by a DFD to get magnitude spectrum and then inverse DCT is applied to compress the data, Mel Scale is applied to distinguish frequencies while finally, a log is applied to the signals to get log energy. Specifically, in the initial stage, the Fourier transform is applied to the voice signals, and the Mel-frequency filter bank is applied subsequently. The above-mentioned filter bank splits the spectrum according to the Mel-scale. It has been found that the lower frequencies have smaller bandwidths in comparison to higher frequencies. Furthermore, frequency spacing of less than 1000 Hz is observed on the Mel-scale, whereas logarithmic spacing is greater than 1000 Hz. In the end, the final section includes a range of coefficients depending on their significance. This is accomplished by calculating the distinct cosine transform of the logarithmic results. The details are shown in Fig 3.2 below.

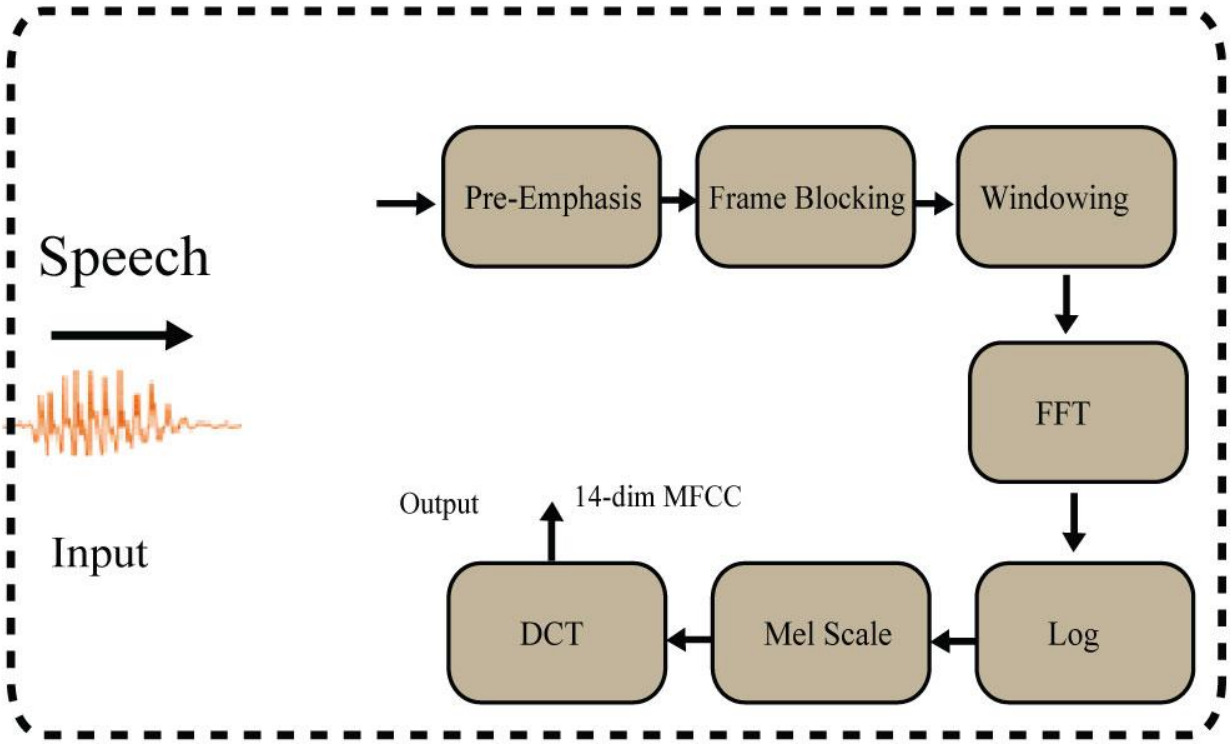

Fig 3.2: MFCC features.

3.5. Gammatone Cepstrum Coefficients

In this section, first the voice input signal is taken and obtained the 14-dim GTCC features for more evaluation. This is also another filter bank-based method developed for obtaining features of cepstral domain. The gammatone feature has several properties that make GTCC filters suitable for modelling human hearing and the system's spectral response. The value of gammatone function is calculated by multiplying the sine tone by the gamma distribution function.

$$gt(t) = Kt^{n-1}e^{-2\pi Bt}\cos(2\pi f_c t + \varphi) \quad t \geq 0 \quad (3.1)$$

Where *K, n, B, fc* and in the above equation (1) denote the amplitude coefficient, filter order, bandwidth parameter, filter center frequency, and phase shift, respectively.

The equal rectangular bandwidth (ERB) is computed by the following equation 2:

$$ERB = \left[\left({}^{fc}/_{EarQ}\right)^{n} + minBW^{n}\right]^{1/n} \quad (3.2)$$

Where *fc*, *EarQ*, *minBW*, and n denotes the center frequency of the filter, the asymptotic quality in the high frequency zone, the minimum bandwidth in the low frequency zone, and the order of the approximation.

The f_{ci} is calculated in equation (3.3).

$$f_{ci} = (f_h + EarQminBW)e^{-\frac{i\,step}{EarQ}} - EarQminBW \quad (3.3)$$

where denotes the higher frequency, *minBW* denotes the ERB parameters, while *i* GT filter index that is given in above equation (3.2). The stage is calculated by equation (3.4).

$$Stage = {EarQ}/{N} \ln\left({f_h + EarQ\ minBW}/{f_l + EarQ\ minBW}\right) \quad (3.4)$$

Where N denotes the number of filters in equation (3.3). The procedure of obtaining features from GTCC is the same as the MFCC, but GTCC uses a gamma tone filter bank and MFCC uses a mel filter bank. 14-dim features were obtained from the voice input signal. The feature extraction of MFCC and GTCC are same, but there is one difference in GTCC, GT filters are applied while in MFCC Mel Filters are applied. In the early stage, windowing of the signals is done followed by FFT for getting magnitude spectrum, which is then mapped to scale, and GT filters are applied, next, log is applied to get log energy and DCT is applied to get a 14- dimensional GTCC features. The information about GTCC feature extraction is given in Figure 3.3 below.

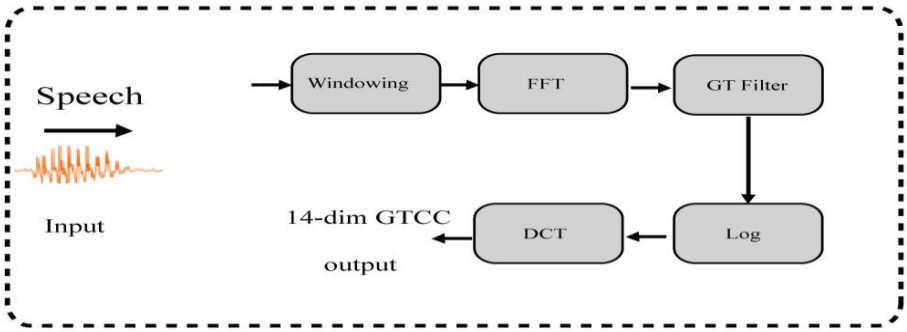

Fig 3.3: GTCC features.

3.6. Spectral Skewness

Next, in this work, spectral skewness feature extracted, which is the measure of the anomaly of dispersal of probability. In this work, a 1-dim feature is extracted from speech, spectral skewness.

The dispersal of skew-normal consists of three parameters, which have curves [27]. The probability density function of dispersal of skew-normal consisting of shape parameter represented by $\alpha \in \mathbb{R}$, which is a scale represented by $\theta \in \mathbb{R}$, while the location is represented by $\in \mathbb{R}$ that is shown in equation (3.5).

$$f(f_h; \xi, \theta, \alpha) = \frac{2}{\theta}\phi\left(\frac{f_h - \xi}{\theta}\right)\boldsymbol{\Phi}\left(\alpha\left(\frac{f_h - \xi}{\theta}\right)\right), f_h \in \mathbb{R}, \tag{3.5}$$

In equation (3.5), ϕ represents a function named normal probability density and this symbol f shows the cumulative distribution function.

Moreover, spectral skewness has three skews, specifically, zero, left, and right. Right is considered as positive and the dispersal of positive skewed is maximum of the right side comparatively on the left side. In addition, negative skew is called left skew and the dispersal of the skew of the left is maximum on the left side comparatively to the right side.

3.7. Spectral Energy

Next, the spectral energy feature of 1-dim is extracted, which is the computation by taking the square root of the mean squared breadth in a short time span. The spectral energy is an expressive approach for computing the average values in a short time span. In the first stage, the breadth of the speech signal is squared, next, it is averaged over a short span of time while finally, the computation of the result is squarely rooted. Specifically, the spectral energy is computed by employing equation 3.6.

$$Spectral\ Energy\ (y) = \sqrt{{}^{1}/_{n}\sum_{n}|y(n)|^2} \tag{3.6}$$

3.8. Corpus

In this research work, the performance of the approach is evaluated by utilizing ASVspoof 2019 LA corpus, which is consisted of three sets, namely, training set, development set, and evaluation. For the training and evaluation, training and evaluation sets are utilized, respectively. The spoof samples of the LA corpus are generated by utilizing 6 cloning spoofing systems, which contain

two Text to speech and four voice conversion systems. The detail of LA corpus is provided in Table 1.

Table 4.1: Detail of LA Corpus.

Logical Access	**Training samples**	**Evaluation samples**	**#Speaker**	
			Male	**Female**
Total samples	15,981	14,161	16	24
Genuine samples	2,580	2,580	8	12
Spoofed samples	13,401	11,581	8	12

3.9. Architecture of Classifier

In this section, the classification performed in this study is discussed. Figure 3.4 demonstrates the classification employed for anti-spoofing measures. Speech input is processed to obtain features and sent to an LSTM network to classify original and fake voices. LSTM has been utilized in many applications. LSTM is a type of RNN used for both time series forecasting and Natural Language Processing (NLP). Acoustic signals are also included in time-series data. The difference between LSTM and Bi-LSTM is the flow of information in both networks. In Bi-LSTM, the flow of information is bi-directional i.e., forward, and backward direction while in the LSTM, the flow of information is unidirectional. Unidirectional flow allows LSTM networks to use peculiar details from the networks. LSTM is also seen as an efficient method for modelling unidirectional dependencies. In summary, LSTM has an additional LSTM layer that alters the flow of information. Specifically, six different layers of LSTM is used, namely, an input layer, which takes an input and further process it to the next layer for computations, three hidden layers that have hidden unit sizes to compute the input and analyse the audio, a SoftMax layer that is used as an activation function, and at the end a classification layer to classify the audio either as spoof or genuine.

The architecture of the LSTM framework applied in this study is shown in Fig 3.4.

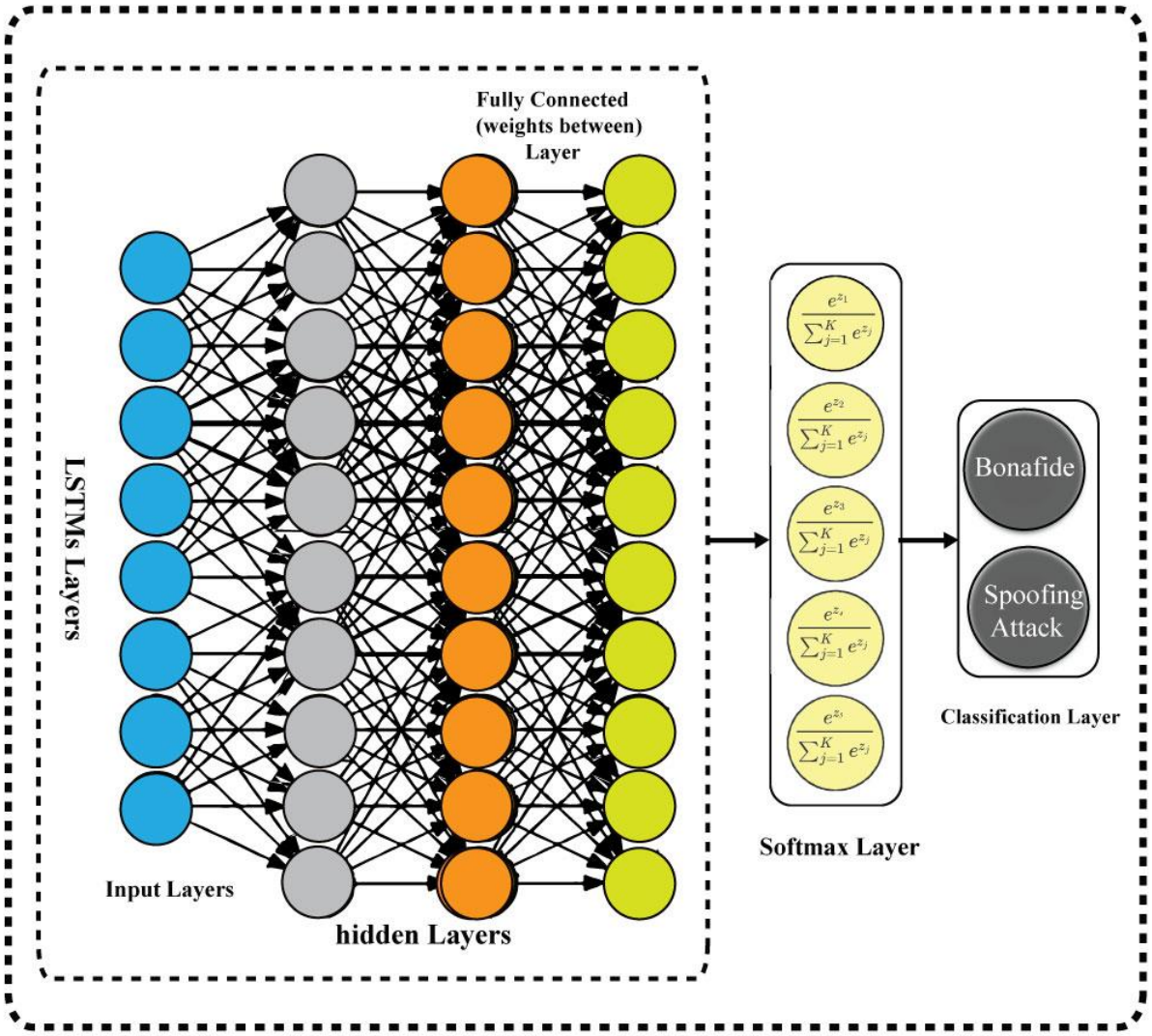

Fig 3.4: Architecture of the LSTM.

3.10. Parameters for Evaluation

The following evaluation parameters will be used for analysis:

- Accuracy.
- Precision.
- Recall.
- F1-Score.

4. Experimental Outcomes

This chapter describes the experimental outcomes in detail. In the below part, the basic system is described, the experimental protocol, the system's performance, its performance using ML algorithms, and its performance compared with the previously developed system. Moreover, the significance of the proposed model is examined with the following evaluation parameters as mentioned in section 3.10 while the comparison is made based on an EER.

4.1. ASVspoof Organizers

The ASVspoof organizers have released three series for the research community. ASVspoof2015 contains data for audio duplication or audio synthesis attacks, ASVspoof2017 contains data for audio duplication attacks, and ASVspoof2019 series contains data for both PA and LA attacks. ASVspoof2019 has two sub-datasets: ASVspoof2019 PA and ASVspoof2019 LA datasets. In the proposed research, the dataset ASVspoof2019 LA was used for performing experiments. All experimental details, including conventional ML and LSTM, are described in the sections that follow.

4.2. Protocols for Experimentation

This section gives information about the experimental code used in this research. The corpus is separated into three parts, i.e., training, development, and evaluation sets. The training set is utilized for the training LSTM model, whereas the evaluation of the LSTM model is performed by using the testing dataset.

4.3. Evaluation of the Approach

This part of the thesis introduces the detailed performance assessment of the proposed system for the identification of audio spoofing attacks. The proposed model is evaluated on the ASVspoof2019 LA dataset. The dataset is categorized into three parts, i.e., training, development, and evaluation sets. This audio from the development set cannot be used for anti-spoofing evaluations. This study utilized the training set for the purpose of training the model and evaluation set for evaluating the trained model, therefore, there is no need to randomly select the training or testing samples. The details of the spoof samples generated for the training and evaluation sets are

given in [28]. Therefore, to achieve the objective of identifying the voice spoofing attacks, audio signals are taken and extracted the 14-dim MFCC, 14-dim GTCC, 1-dim spectral skewness, and 1-dim spectral energy features from both sets, especially the training and evaluation set of ASVspoof2019 LA. The details about the feature extraction and classification are explained in 3rd chapter. Several DL models demonstrated high classification efficiency on time series data. Moreover, voice is time-series data, and the LSTM architecture gives astonishing results. Information about the results of anti-spoofing measures is shown in Figure 4.1. The proposed method achieved an exceptional 100% accuracy in the binary classification of authentic and fake audio samples. The precision results of about 99.88% proved that the proposed model is efficient for the identification of fake voice signals. The system achieved a recall of 100%, along with an F1-score of 99.94%. The constant CQCC and GMM were used as classifiers in the ASVspoof organizer baseline. In addition, authors utilized Linear Cepstrum Coefficients (LFCC) for baseline and classification performed using GMM. Although, the resulting system is unreliable for use in a real-time environment as its capabilities do not gather enough details. However, the EER value of 0.01% is significantly lower compared to baseline techniques [38]. Baseline speech recognition approaches achieved EERs of 11.04% and 13.54% by utilizing the two approaches, namely, CQCC-GMM and LFCC-GMM, separately. This approach yielded 11.03% and 13.53% lower EER values compared to these approaches [28]. The corpus, specifically ASVspoof2019 LA, contains speech samples generated with several powerful algorithms. The size used for generating spoofing attacks also varies in size such as 2-5m, 5-10m, and 10-20m ranges. Although the utilized datasets are diverse, the developed method achieves 100% accuracy, specifying that the system is effective and reliable in identifying voice spoofing attacks.

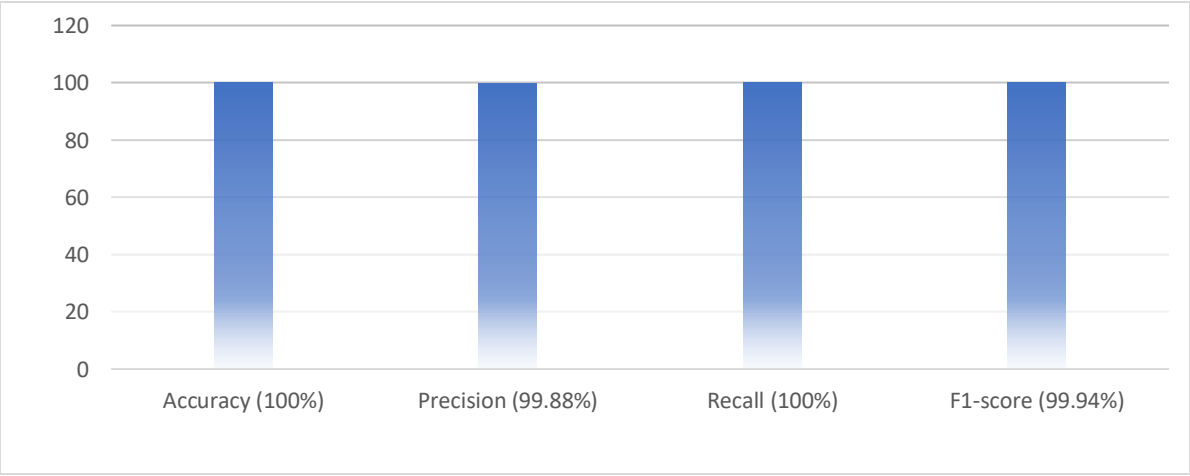

Fig 4.1: Performance of the Approach.

4.4. Error Matrix of Proposed Approach

The error matrix contains information about the examination of the system's categorization outcome as detailed in Table 4.2. Confusion matrices are formed specifically for the problems of categorization and specifically have 4 evaluates True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). TP indicates correct forecasts for the positive class and TN indicates correct forecasts for the negative class. Likewise, FP indicates false forecasts for the positive class and FN indicates false forecasts for the negative class. The values for TP, FP, FN, and TN are 2580, 3, 0, and 15900, respectively, as shown in the Table 4.2 confusion matrix prepared for the system. The values achieved above show that the developed system correctly classifies all authentic audios and not a single duplicate voice is identified as an authentic voice. Likewise, the system recognized 15900 spoofed voices as spoofed, but only 3 spoofed voices were recognized as authentic. Overall, 0% of the voices were misclassified, and the remaining 100% of the voices were accurately recognized. In the confusion matrix, 1 represents the authentic class and 2 represents the spoofing class.

Table 4.2: Confusion Matrix.

Predicted Class	**Actual Class**		
		Genuine	Spoof
	Genuine	2580	3
	Spoof	0	15900

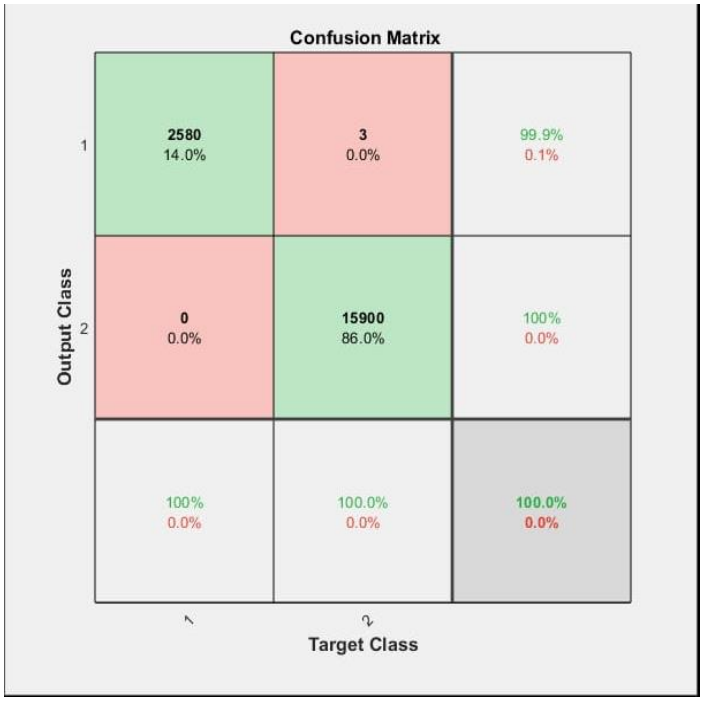

Fig 4.2: Confusion Matrix in Percentage.

4.5. Accuracy and Loss of Model

The accuracy and loss of model graphs of this approach are presented in Fig 4.3.

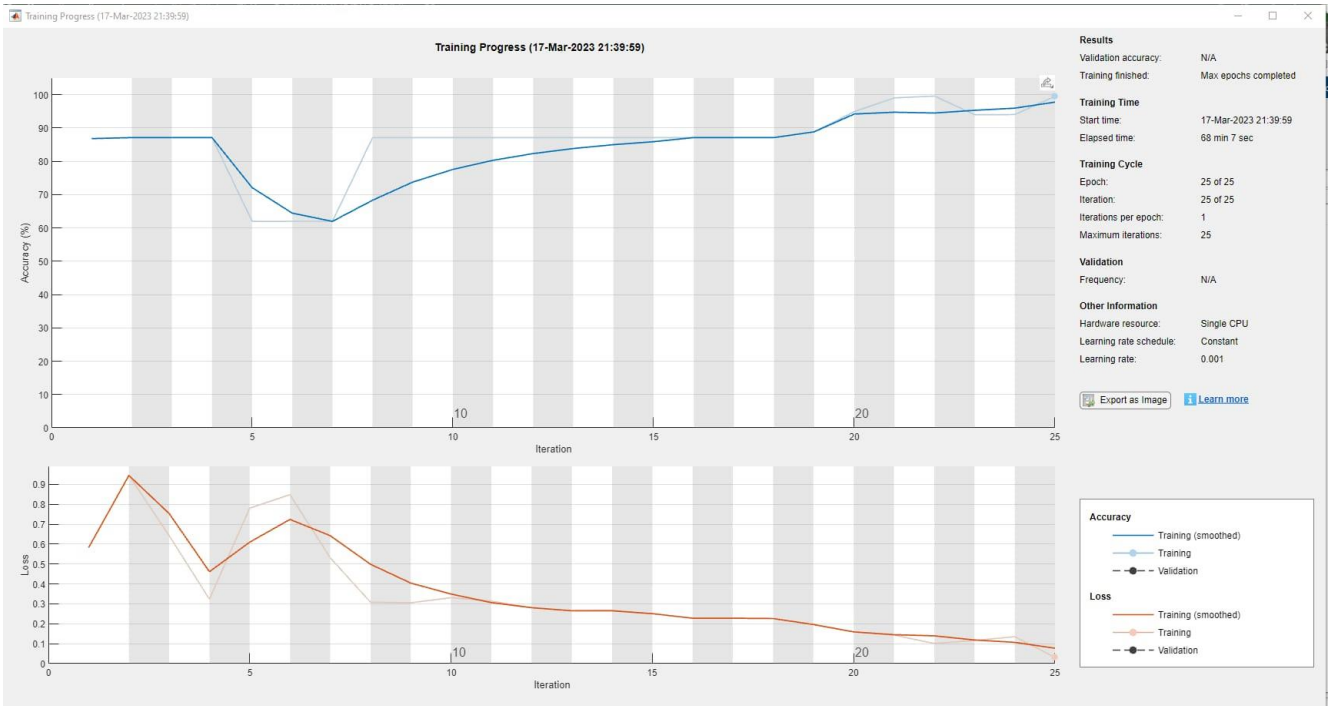

Fig 4.3: Model Accuracy Graph.

4.6. Evaluation of Conventional Algorithms

This section comprehensively explained traditional algorithms for identifying spoofed and authentic speech. The performance of SVMs, ensembles, ANNs, decision trees, discriminant analysis, and Naive Bayes has been examined. The results are discussed in detail.

4.6.1. Evaluation of the SVM

In this section, the significance of SVM is evaluated for identifying audio spoofing attacks. Various applications have used the SVM for classification purposes. First, 14 and 14-dim features of GTCC and MFCC, and 1, 1-dim of spectral energy and spectral skewness were obtained for training the classifier, SVM. In addition, liner kernel is used. Fig 4.4 depicts the detailed outcomes of SVM. As shown in Figure 4.4, SVM achieved 78.03% accuracy, which is much lower than the 100% accuracy, that the approach has obtained. The recall of SVM is 81.53% while the F1-score is 74.31%, separately.

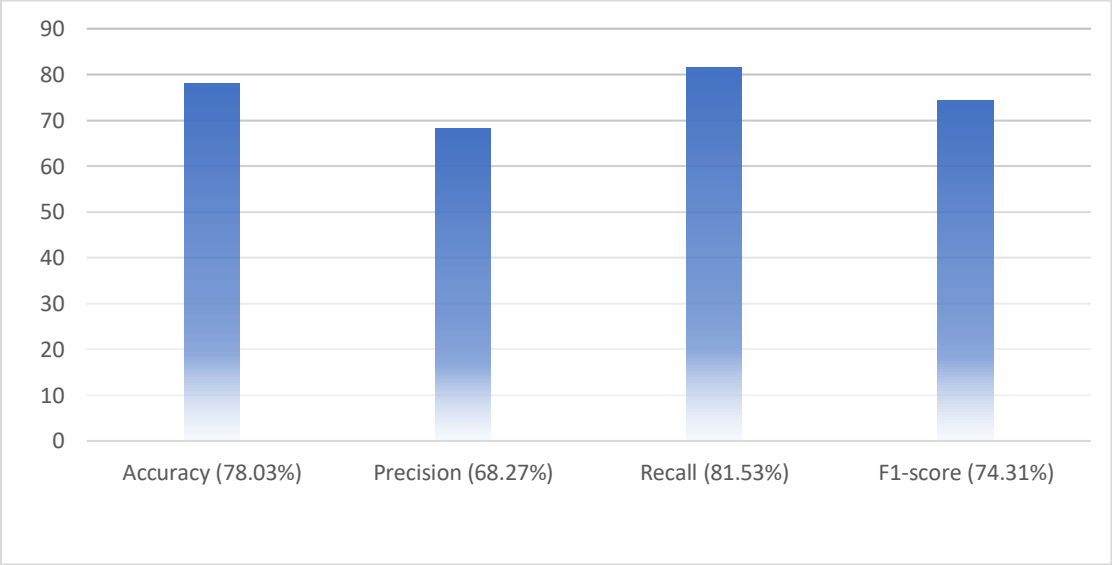

Fig 4.4: Performance of the SVM.

4.6.2. Error-Matrix of the SVM

An error matrix is developed to observe the efficiency of identifying voice spoofing attacks for the SVM classifier. The four values, specifically TP, FP, FN, and TN, are detailed in Fig 4.5.

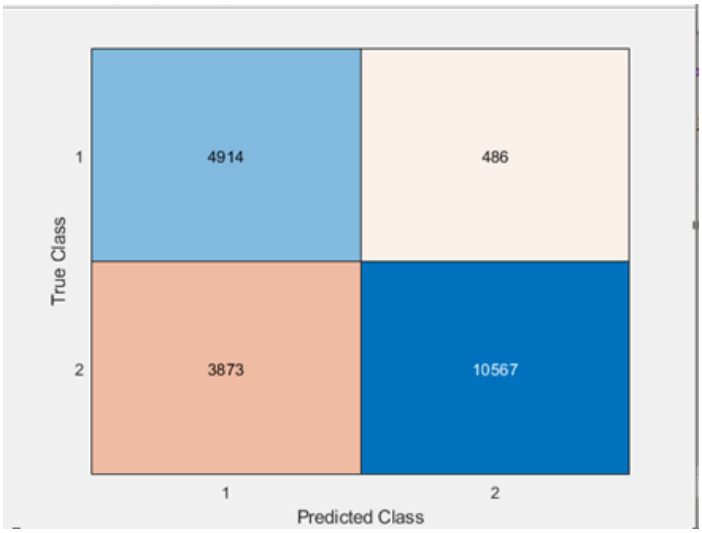

Fig 4.5: Confusion Matrix of the SVM.

4.6.3. Performance of the Ensemble Classifier

The information about the ensembled classifier is given in Figure 4.6. In addition, these parameters were utilized, namely, maximum number of split equals 5, robust boost, and number of bins equals 5, respectively. An ensemble classifier has yielded an accuracy of 84.71%, which compared to the SVM is better. Likewise, the recall of an ensemble is 81.53% while the F1-score is 76.63%, separately.

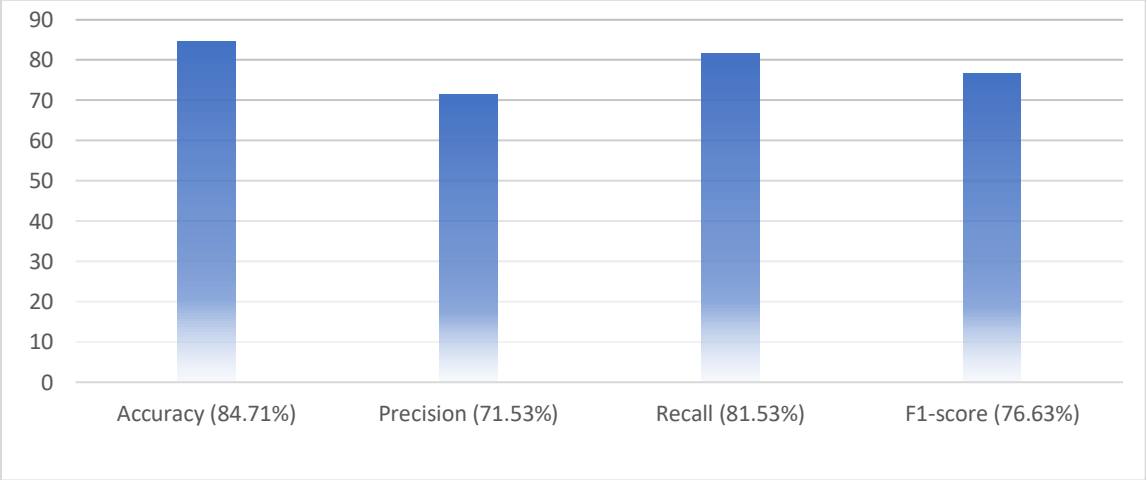

Fig 4.6: Performance of the Ensemble Classifier.

4.6.4. Error Matrix for Ensemble Classifier

Fig 4.7 shows comprehensive categorization performance results for TP, FP, FN, and TN values. In Fig 4.7, it is confirmed that the ensemble classifier correctly classified 3644 and 13162 authentic and fake voices and misclassified 1756 and 1278 voice samples.

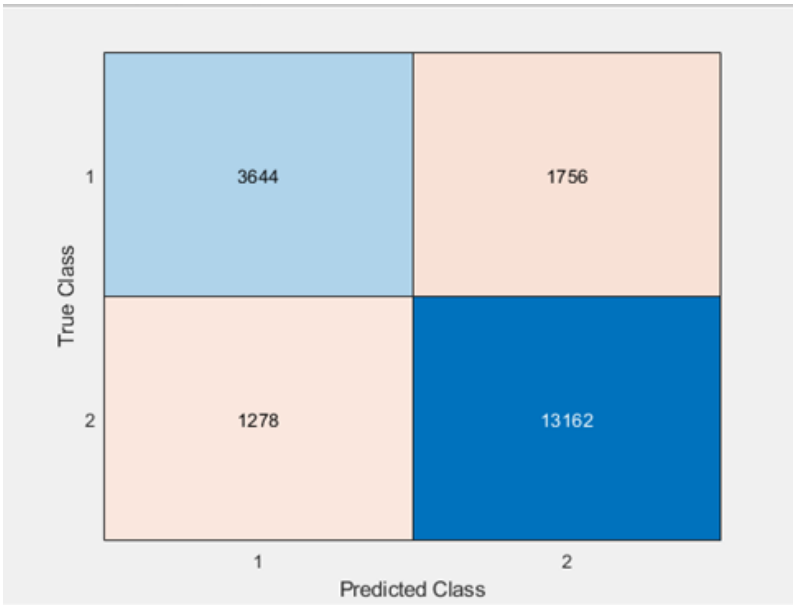

Fig 4.7: Confusion Matrix of the Ensemble Classifier.

4.6.5. Evaluation of the KNN Classifier

K-nearest neighbor (KNN) classifiers have been used in a variety of applications. In addition, I used 15 neighbors, distance is cosine, and cost equals to [0,3: 1,0] A total of 30-dim features is obtained and provided to KNN to categorize authentic and fake voices.

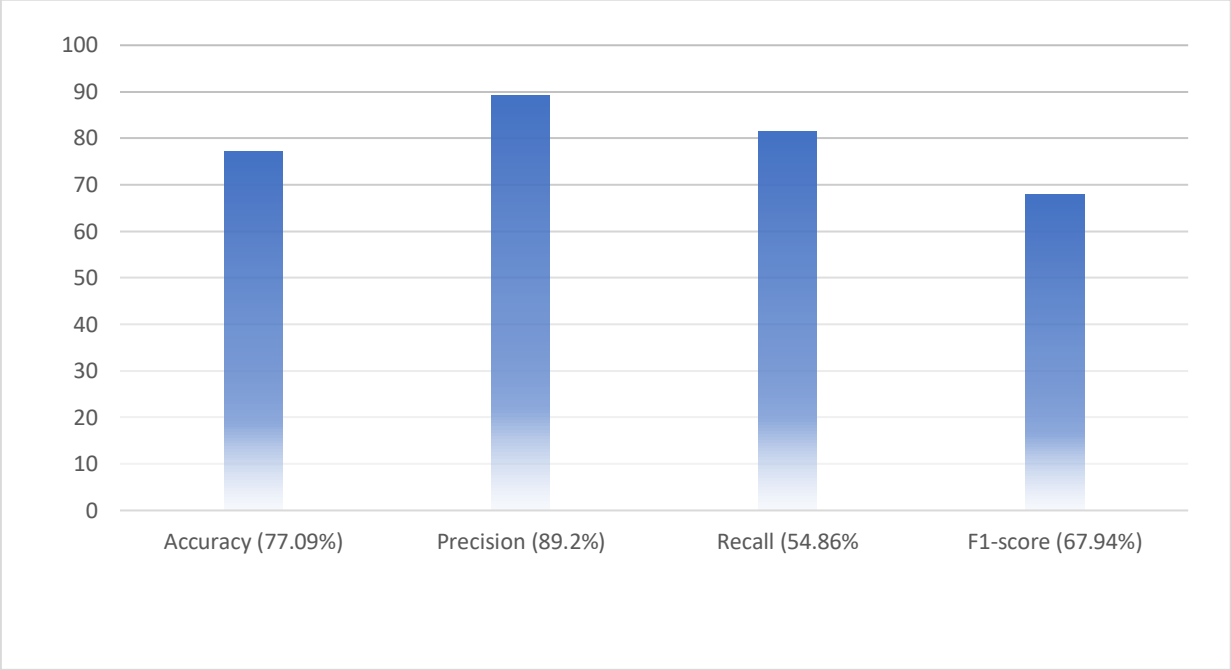

Fig 4.8: Performance of the KNN Classifier.

4.6.6. Error-Matrix of the KNN Classifier

The detailed experimental outcomes in terms of TP, FP, FN, and TN are depicted in Fig 4.9 for the KNN. As it can be checked, the TP, FP, FN, and TN values are 4817, 583, 3963, and 10477, separately. The correct and incorrect classification outcomes are shown in Fig 4.9.

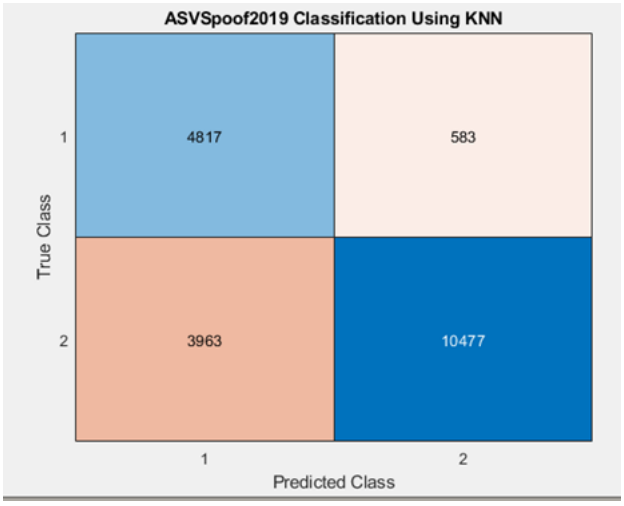

Fig 4.9: Confusion Matrix of the KNN Classifier.

4.7. Evaluation Comparison with Existing Systems

This method uses integration of mel frequency cepstral coefficients (MFCC), gamma tone cepstral coefficients (GTCC), spectral energy, and spectral skewness while for the purpose of classification, long short-term memory network (LSTM) is employed. Next, the performance of the approach with other approaches is compared as shown in Fig 4.10. The experiments are performed using MFCC-LSTM, GTCC-LSTM, and MFCC-GTCC-LSTM; however, the performance is poor, therefore, a novel integration of spectral features MFCC, GTCC, spectral energy, and spectral skewness is used, and the performance significantly improved.

This experiment has details of the performance comparison of this approach against the existing approaches [28-33]. As shown in Fig 4.10, the performance of this approach is compared with the baseline [28] as well as other approaches for demonstrating the superiority of the proposed mechanism. The EER values of the existing approaches [38-43] and proposed approach are reported in Fig 4.10. From the outcomes, it is noticed that the approach has the best performance compared to the baseline [28] and existing approaches [28-33]. In terms of second-best performance, the approach [43] has obtained an EER of 4.53% while the baseline approach (CQCC-GMM as well as LFCC-GMM) [28] has the poorest performance and has an EER of 9.57% and 8.09%, respectively. Observing the comprehensive outcomes as reported in Fig 4.10, there is a significant reduction in the EER value by 9.56%, 8.08%, 7.65%, 6.37%, 6.27%, 5.31%, and 4.52%, respectively than these approaches [28-33]. The above significant reduction in the EER value of this approach compared to the existing approaches shows that the proposed working mechanism has superior performance in the detection of voice spoofing attacks. In addition, comparative assessment with conventional algorithms and experimental outcomes confirm that the proposed working mechanism has a smaller EER value and outperforms existing approaches. Specifically, the above discussion confirms that the proposed approach is reliable to be utilized in ASV systems for the detection of LA attacks. Detailed comparative analysis is provided in Fig 4.10 with existing approaches.

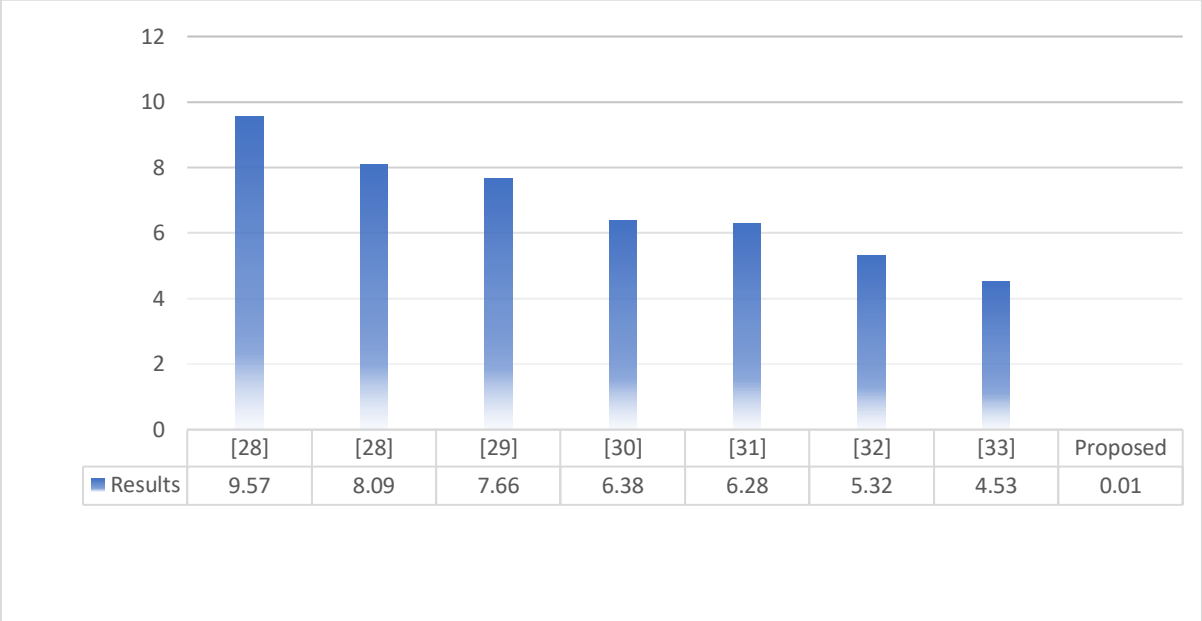

Fig 4.10: Comparative Analysis with Existing Systems.

5. Conclusion

5.2. Conclusion

This study proposed a new method for identifying spoofing attacks. Hackers use powerful algorithms to generate the voices of authentic users and play against ASV to gain unauthorized approaches to systems for unethical aims. This kind of spoofing attack is considered a big threat to the security of automated speech identification systems. To increase security and protect automated speech identification systems from voice spoofing attacks, this study developed a novel technique based on the integration of features. In this study, MFCC, GTCC, spectral energy, and skewness are used as feature extraction techniques, and LSTM is used for classifying authentic and fake voices. Additionally, the ASVspoof2019 LA dataset is used for the experimental process. The proposed approach achieves 100% accuracy with a precision rate of 99.88%. The recall and F1 score values are 100% and 99.94%, respectively.

5.2. Future Work

In the future, the main aim is to employ similar approaches for voice replay attacks. Although the proposed approach has significant experimental outcomes against LA attacks, however, the effectiveness of this approach has not been investigated against PA attacks and deep fake attacks. Therefore, the performance of this approach against PA and deep fake attacks can be checked in future.

REFERENCES

[1] Scardapane, Simone, et al. "On the use of deep recurrent neural networks for detecting audio spoofing attacks." *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017.

[2] Sriskandaraja K, Sethu V, Ambikairajah E. Deep siamese architecture-based replay detection for secure voice biometric. InInterspeech 2018 Sep (pp. 671-675).

[3] Chen, Zhuxin, et al. "Recurrent neural networks for automatic replay spoofing attack detection." *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018.

[4] Huang, Lian, and Chi-Man Pun. "Audio replay spoof attack detection using segment-based hybrid feature and densenet-LSTM network." *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2019.

[5] Tian, Xiaohai, et al. "An Investigation of Spoofing Speech Detection Under Additive Noise and Reverberant Conditions." *INTERSPEECH*. 2016.

[6] Hanilci, Cemal, et al. "Spoofing detection goes noisy: An analysis of synthetic speech detection in the presence of additive noise." *Speech Communication* 85 (2016): 83-97.

[7] Bharath, K.P. and Kumar, M.R., 2022. New replay attack detection using iterative adaptive inverse filtering and high frequency band. *Expert Systems with Applications*, *195*, p.116597.

[8] Patil, Ankur T., et al. "Improving the potential of enhanced teager energy cepstral coefficients (etecc) for replay attack detection." *Computer Speech & Language* 72 (2022): 101281.

[9] Gunendradasan, Tharshini, et al. "An adaptive transmission line cochlear model based front-end for replay attack detection." *Speech Communication* 132 (2021): 114-122.

[10] Aljasem, Muteb, et al. "Secure automatic speaker verification (sasv) system through sm-altp features and asymmetric bagging." *IEEE Transactions on Information Forensics and Security* 16 (2021): 3524-3537.

[11] Nasersharif, Babak, and Morteza Yazdani. "Evolutionary fusion of classifiers trained on linear prediction-based features for replay attack detection." *Expert Systems* 38.3 (2021): e12670.

[12] Yue, Ling, et al. "LiveEar: An Efficient and Easy-to-use Liveness Detection System for Voice Assistants." *Journal of Physics: Conference Series*. Vol. 1871. No. 1. IOP Publishing, 2021.

[13] Javed, Ali, et al. "Towards protecting cyber-physical and IoT systems from single-and multi-order voice spoofing attacks." *Applied Acoustics* 183 (2021): 108283.

[14] Yaguchi, Ryoya, et al. "Replay attack detection based on spatial and spectral features of stereo signal." *Journal of Information Processing* 29 (2021): 275-282.

[15] Wei, Linqiang, et al. "New acoustic features for synthetic and replay spoofing attack detection." *Symmetry* 14.2 (2022): 274.

[16] Prajapati, Gauri P., Madhu R. Kamble, and Hemant A. Patil. "Energy separation-based features for replay spoof detection for voice assistant." *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021.

[17] Khoria, Kuldeep, Ankur T. Patil, and Hemant A. Patil. "Significance of constant q transform for voice liveness detection." *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021.

[18] Dutta, Krishna, Madhusudan Singh, and Debadatta Pati. "Detection of replay signals using excitation source and shifted cqcc features." *International Journal of Speech Technology* 24 (2021): 497-507.

[19] Meng, Yan, et al. "Your microphone array retains your identity: A robust voice liveness detection system for smart speakers." *31st USENIX Security Symposium (USENIX Security 22)*. 2022.

[20] Mittal, Aakshi, and Mohit Dua. "Static–dynamic features and hybrid deep learning models-based spoof detection system for ASV." *Complex & Intelligent Systems* 8, no. 2 (2022): 1153-1166.

[21] Ren, Yanzhen, et al. "Replay attack detection based on distortion by loudspeaker for voice authentication." *Multimedia Tools and Applications* 78 (2019): 8383-8396.

[22] Yoon, Sung-Hyun, et al. "A new replay attack against automatic speaker verification systems." *IEEE Access* 8 (2020): 36080-36088.

[23] Garg, Sachin, Shruti Bhilare, and Vivek Kanhangad. "Subband analysis for performance improvement of replay attack detection in speaker verification systems." *2019 IEEE 5th International Conference on Identity, Security, and Behavior Analysis (ISBA)*. IEEE, 2019.

[24] Gunendradasan, Tharshini, et al. "Transmission line cochlear model-based AM-FM features for replay attack detection." *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

[25] Singh, Madhusudan, and Debadatta Pati. "Usefulness of linear prediction residual for replay attack detection." *AEU-International Journal of Electronics and Communications* 110 (2019): 152837.

[26] Tak, Hemlata, and Hemant A. Patil. "Novel Linear Frequency Residual Cepstral Features for Replay Attack Detection." *INTERSPEECH*. 2018.

[27] Azzalini, Adelchi. "The skew-normal distribution and related multivariate families." *Scandinavian journal of statistics* 32.2 (2005): 159-188.

[28] Todisco, Massimiliano, et al. "ASVspoof 2019: Future horizons in spoofed and fake audio detection." *arXiv preprint arXiv:1904.05441* (2019).

[29] Chettri, Bhusan, et al. "Ensemble models for spoofing detection in automatic speaker verification." *arXiv preprint arXiv:1904.04589* (2019).

[30] Zhang, Chunlei, Chengzhu Yu, and John HL Hansen. "An investigation of deep-learning frameworks for speaker verification antispoofing." *IEEE Journal of Selected Topics in Signal Processing* 11.4 (2017): 684-694.

[31] Gomez-Alanis, Alejandro, et al. "A light convolutional GRU-RNN deep feature extractor for ASV spoofing detection." *Proc. Interspeech*. Vol. 2019. 2019.

[32] Aravind, P. R., Usamath Nechiyil, and Nandakumar Paramparambath. "Audio spoofing verification using deep convolutional neural networks by transfer learning." *arXiv preprint arXiv:2008.03464* (2020).

[33] Lavrentyeva, Galina, et al. "STC antispoofing systems for the ASVspoof2019 challenge." *arXiv preprint arXiv:1904.05576* (2019).