A dissertation submitted to the **University of Greenwich**
in partial fulfilment of the requirements for the Degree of

Master of Science

in

Data Science

Fake Voice Detection using Machine Learning

Name: Kruthika Mysore Bhaskar

Student ID: 0013554599

Supervisor: Seyed Jazdarreh

Submission Date: 8TH September 2025

Word Count: 14235

FAKE VOICE DETECTION USING MACHINE LEARNING

Computing & Mathematical Sciences, University of Greenwich,
30 Park Row, Greenwich, UK.

(Submitted 8th September 2025)

ABSTRACT

Recent developments in speech synthesis and voice conversion have transformed communication, accessibility and entertainment. While these technologies provide valuable benefits, they also introduce serious security concerns. Artificially generated voices, often referred to as deepfakes, are now able to mimic human speakers with striking accuracy. Such capabilities can be misused in fraud, misinformation and identity-related crimes, making the detection of synthetic voices a pressing challenge within biometric security and automatic speaker verification (ASV).

This dissertation investigates methods for detecting spoofed audio using the ASVspoof 2019 Logical Access (LA) dataset. The research focuses on the use of Linear Frequency Cepstral Coefficients (LFCCs) for feature extraction. Unlike the widely applied Mel Frequency Cepstral Coefficients (MFCCs), LFCCs retain finer spectral details, allowing them to capture the subtle distortions introduced by voice conversion and synthesis techniques. These features were used to train Convolutional Neural Networks (CNNs), tested in both a baseline configuration and with class weighting to address class imbalance within the dataset.

The performance of the system was evaluated using Equal Error Rate (EER) as the primary measure, supported by accuracy, precision, recall and F1-score. The CNN achieved an EER of 0.20 percent, reflecting a strong ability to distinguish between genuine and spoofed voices. Accuracy was close to 100 percent, while precision, recall and F1-scores also produced near-perfect results. These outcomes demonstrate the potential of LFCC-based CNNs for anti-spoofing tasks, while also highlighting the need for further investigation into their resilience when faced with unfamiliar spoofing techniques and practical deployment scenarios.

The study makes a contribution to ongoing work in speech security by providing a detailed evaluation of LFCC-driven CNN models. It shows the effectiveness of LFCCs in exposing spoofing artefacts, stresses the importance of Equal Error Rate as an evaluation metric and identifies future research opportunities aimed at developing more reliable and adaptable voice detection systems

PREFACE

This dissertation was undertaken as part of the requirements for the Master of Science in Data Science at the University of Greenwich. The selection of this project reflects both an academic commitment and a genuine interest in the growing challenges created by artificial intelligence in daily life. As synthetic voice and deepfake technologies continue to advance and produce increasingly convincing results, research in this area has become essential for protecting digital trust and security.

The project provided an opportunity to integrate knowledge of data science, machine learning and cybersecurity in addressing a problem with clear practical relevance. By working with the ASVspoof 2019 dataset, the study engaged with a recognised benchmark in the field. The exploration of Linear Frequency Cepstral Coefficients (LFCCs) alongside Convolutional Neural Networks (CNNs) allowed for an examination of contemporary techniques in the detection of manipulated audio and contributed to the wider field of voice anti-spoofing.

Beyond fulfilling academic requirements, the dissertation reflects an ongoing curiosity about the responsible and secure use of technology. It is intended not only as a formal academic study but also as a modest contribution to broader efforts aimed at countering the misuse of artificial intelligence in audio generation.

ACKNOWLEDGEMENTS

I would like to sincerely thank **Professor Seyed Jazdarreh** for his invaluable guidance, encouragement, and constructive feedback throughout my project. His expertise and support have been instrumental in shaping my research and helping me bring it to completion.

I am also grateful to the faculty and staff of the **University of Greenwich** for providing the academic environment, resources, and technical support that enabled me to carry out this work. I would further like to acknowledge the ASVspoof 2019 Challenge organizers for making their dataset publicly available, which formed the foundation of my study.

Finally, I would like to express my heartfelt thanks to my family, friends, and family friends for their patience, understanding, and constant encouragement. Their unwavering support has been a source of strength throughout this journey and has made this achievement truly meaningful.

Table of Contents

Table of Figure

1.INTRODUCTION

1.1 Overview

In recent years, speech technologies have advanced at an exceptional pace. Deep learning models such as WaveNet and Tacotron have made it possible to generate voices that are strikingly human in their rhythm, tone and clarity (van den Oord et al., 2016; Shen et al., 2018). These innovations have had a significant positive impact, particularly in areas such as accessibility for individuals with disabilities, the development of digital assistants and applications in entertainment.

At the same time, the very technologies that have enabled these advances have also created new risks. Synthetic voices are now increasingly misused for malicious purposes, including fraud, impersonation and the spread of misinformation. High-profile incidents have already shown how cloned voices have deceived employees into transferring large sums of money, illustrating the seriousness of this threat (Kietzmann et al., 2020). Beyond financial crime, synthetic audio also poses dangers in the manipulation of public opinion and the undermining of audio evidence in legal or forensic settings (Korshunov and Marcel, 2018). The result is a growing erosion of trust in voice, which has traditionally been regarded as a secure biometric and a reliable medium of communication. This erosion has made the detection of fake voices a critical area of research.

Automatic Speaker Verification (ASV) systems, which are widely deployed for authentication, are particularly vulnerable to such attacks. They can be compromised through replayed audio, text-to-speech (TTS) synthesis or voice conversion (VC) methods (Todisco et al., 2019). To support research addressing these vulnerabilities, the ASVspoof Challenge was introduced, providing standardised benchmarks and datasets. The ASVspoof 2019 Logical Access (LA) dataset, used in this study, contains both genuine and spoofed samples, thereby offering a realistic and controlled environment for the evaluation of anti-spoofing systems.

This dissertation applies Linear Frequency Cepstral Coefficients (LFCCs) to extract acoustic features from the audio samples. In contrast to the more widely used Mel Frequency Cepstral Coefficients (MFCCs), LFCCs employ a linear frequency scale, which makes them more effective in detecting high-frequency artefacts that often arise during speech synthesis and conversion

(Sahidullah, Kinnunen and Hanilçi, 2015). These features are then classified using a Convolutional Neural Network (CNN), which is trained to distinguish between genuine and spoofed speech.

Performance is primarily evaluated using Equal Error Rate (EER), a balanced measure that reflects both false acceptances and false rejections. Additional metrics, including accuracy, precision, recall and F1-score, are also reported to provide a comprehensive view of the system's performance (Sahidullah, Kinnunen and Hanilçi, 2015; Todisco et al., 2019).

In summary, this research is situated within the expanding field of voice anti-spoofing. By combining LFCC-based feature extraction with CNN classification and benchmarking against the ASVspoof 2019 dataset, the study seeks to contribute towards the development of more secure and trustworthy speaker verification systems capable of resisting contemporary spoofing attacks.

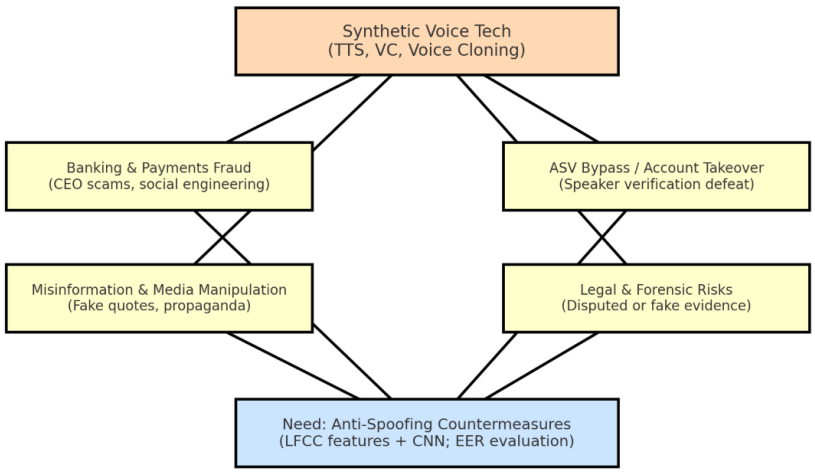

Figure 1 Threat Landscape of voice spoofing attacks.

1.2 Objectives

The central aim of this project is to design and evaluate a machine learning system capable of distinguishing genuine voices from artificially generated ones. To address this aim, the research focuses on the use of Linear Frequency Cepstral Coefficients (LFCCs) as the primary feature representation, combined with a Convolutional Neural Network (CNN) as the classification model. The study is conducted within the framework of the ASVspoof 2019 Logical Access (LA) dataset, which serves as a widely recognised benchmark for the study of synthetic and converted speech (Wang et al., 2020).

In order to achieve this aim, the project is guided by the following objectives:

1. To review and critically analyse existing literature on fake voice detection. This includes an examination of feature extraction techniques such as Mel Frequency Cepstral Coefficients (MFCCs), LFCCs, and spectrogram-based representations, as well as classification methods ranging from Gaussian Mixture Models (GMMs) and Support Vector Machines (SVMs) to more recent deep learning approaches, including CNNs and Recurrent Neural Networks (Sahidullah, Kinnunen and Hanilçi, 2015; Todisco et al., 2019).

2. To prepare and extract LFCC features from the ASVspoof 2019 LA dataset. This process ensures that both genuine and spoofed audio samples are represented in a way that captures the subtle distortions introduced by modern text-to-speech (TTS) and voice conversion (VC) algorithms (Wang et al., 2020).

3. To develop and train a CNN classifier capable of identifying discriminative patterns within LFCC features, with experiments conducted using both baseline training and weighted configurations to address the issue of class imbalance.

4. To evaluate system performance using Equal Error Rate (EER) as the principal measure of effectiveness, supplemented by accuracy, precision, recall and F1-score to provide a more comprehensive assessment of model behaviour (Sahidullah, Kinnunen and Hanilci, 2015).

5. To interpret results and identify limitations, particularly the challenges associated with generalising performance to previously unseen spoofing methods. On the basis of these findings, the study proposes recommendations for future research, including the adoption of advanced deep learning architectures such as ResNets and Transformers, ensemble methods, and the potential development of real-time detection systems (Todisco et al., 2019).

By addressing these objectives, the project seeks to demonstrate the effectiveness of LFCC-based CNN models in detecting synthetic speech. At the same time, it contributes to the wider body of research on voice anti-spoofing by highlighting both the strengths and the limitations of current methods and identifying avenues for future improvement.

1.3 Scope of the Study

The scope of this research is defined by the use of the ASVspoof 2019 Logical Access (LA) dataset, which contains a wide range of both genuine and spoofed audio recordings. This dataset was specifically developed to support the study of synthetic speech and voice conversion attacks and has since become an established benchmark within the anti-spoofing research community (Wang et al., 2020). By adopting this dataset, the study ensures that its findings can be fairly compared with related work and that the experiments are conducted within realistic and widely recognised conditions.

The investigation centres on the application of Linear Frequency Cepstral Coefficients (LFCCs) as the primary feature representation. LFCCs were chosen because they preserve fine spectral detail, particularly in the higher frequency regions where synthesis artefacts are most apparent. Prior research has shown that LFCCs are more effective than alternative features, such as Mel Frequency Cepstral Coefficients (MFCCs), in detecting synthetic speech (Sahidullah, Kinnunen and Hanilçi, 2015). In this study, LFCC features are used as inputs to a Convolutional Neural Network (CNN), which is well suited to learning complex patterns in audio data and has demonstrated strong performance in voice anti-spoofing tasks (Todisco et al., 2019).

The scope of the work is deliberately restricted to the Logical Access condition, which focuses on attacks generated by advanced text-to-speech and voice conversion methods. Other forms of spoofing, such as replay attacks carried out in physical environments, are outside the boundaries of this research. Similarly, the project does not seek to deliver a fully deployable or industry-ready system. Instead, it is designed to provide insight into the behaviour, strengths and limitations of LFCC-driven CNNs when tested under controlled experimental settings.

By narrowing its scope in this way, the research offers a clearer understanding of how LFCC features and CNN architectures perform against synthetic and converted voice attacks. At the same time, it establishes a foundation for future studies that may extend these approaches to broader contexts, including replay-based spoofing, real-time detection systems or multimodal biometric security frameworks.

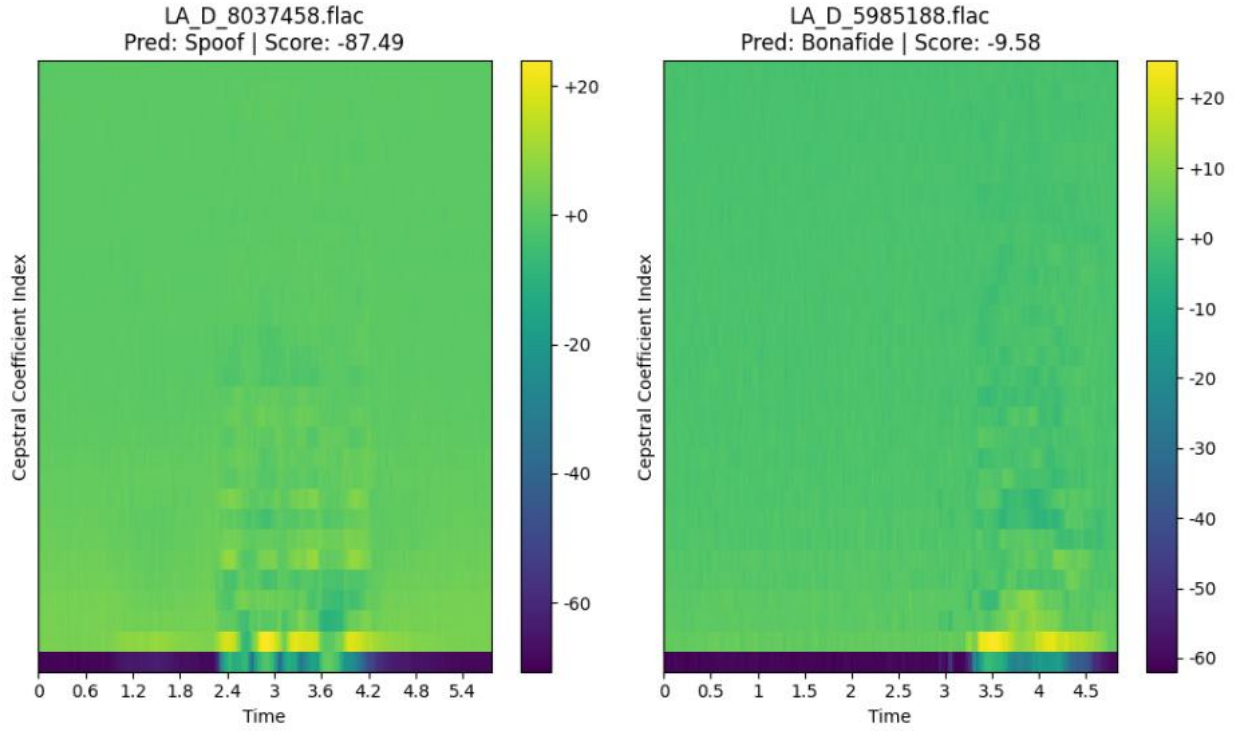

Figure 2 Comparison of genuine vs spoofed audio spectrograms

Spectrograms generated from the ASVspoof 2019 LA dataset. The left image represents a spoofed sample, where irregular spectral artefacts are visible. The right image represents a bonafide

(genuine) sample, showing more stable cepstral patterns. These visual differences highlight the type of spectral cues that LFCC features and CNN models aim to capture for classification.

1.4 Motivation

The motivation for this research is rooted in the increasing misuse of synthetic speech technologies and the urgent requirement for effective countermeasures. Text-to-speech (TTS) and voice conversion (VC) systems have delivered valuable benefits in areas such as accessibility, healthcare, entertainment and customer interaction. Nevertheless, the same technologies are now being exploited for malicious purposes. One widely cited case involved the use of a cloned voice to impersonate a company executive and authorise the fraudulent transfer of funds, demonstrating the real-world risks associated with deepfake audio (Kietzmann et al., 2020). Such incidents illustrate the growing threat of spoofed speech and emphasise the importance of strengthening security in systems that continue to rely on voice as a trusted channel of communication.

Traditionally, voice has been regarded as a reliable biometric identifier. Automatic Speaker Verification (ASV) systems are extensively deployed for applications including banking authentication, forensic investigation and access control. However, recent studies reveal that these systems can be deceived by increasingly sophisticated spoofing techniques, including high-quality synthetic voices (Korshunov and Marcel, 2018; Todisco et al., 2019). These vulnerabilities extend beyond financial risk, undermining the credibility of voice evidence in legal and forensic contexts and raising wider concerns about digital trust.

The research is also motivated by the opportunity to contribute to a rapidly growing body of academic work in this area. Initiatives such as the ASVspoof challenges have played a key role in advancing the field by providing standardised datasets and shared evaluation frameworks (Wang et al., 2020). By employing the ASVspoof 2019 Logical Access dataset, this study situates itself within internationally recognised research practices and ensures that its results can be compared meaningfully with related studies.

From a technical standpoint, the study is driven by the potential of Linear Frequency Cepstral Coefficients (LFCCs) to detect subtle artefacts of synthetic speech that are often overlooked by

traditional features such as Mel Frequency Cepstral Coefficients (MFCCs) (Sahidullah, Kinnunen and Hanilçi, 2015). When combined with the ability of Convolutional Neural Networks (CNNs) to learn complex patterns, this approach offers strong potential to improve both the accuracy and the robustness of anti-spoofing systems.

In summary, the motivation for this project reflects both societal and academic imperatives: to safeguard individuals and organisations against fraud and misinformation, and to contribute to the advancement of knowledge in the field of voice anti-spoofing research.

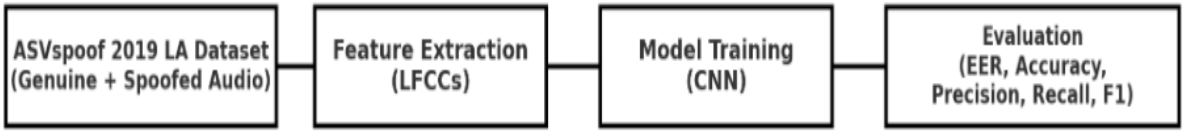

Figure 3 Project Pipeline

dataset -> LFCC feature extraction -> CNN training -> valuation using EER and complementary metrics.

1.5 Challenges

The development of a reliable system for detecting fake voices presents a number of challenges, both technical and conceptual. Overcoming these issues is crucial to ensure that proposed countermeasures are not only effective in controlled settings but also scalable and adaptable to new forms of attack as they emerge.

Class Imbalance

A recurring difficulty in machine learning is the imbalance between classes within a dataset. In many cases, the number of genuine and spoofed samples is uneven, which can bias models towards the majority class and undermine their ability to correctly classify the minority class (Delgado et al., 2014). This issue is present in the ASVspoof 2019 dataset, where strategies such as class

weighting or data augmentation are required to achieve more balanced training and improve performance.

Diversity of Spoofing Attacks

Synthetic speech can be generated using a range of text-to-speech (TTS) and voice conversion (VC) methods, each producing distinct acoustic artefacts. A system trained on one type of spoofing method may fail when exposed to another, particularly when the attack is produced by more advanced or previously unseen algorithms (Todisco et al., 2019). This diversity of approaches makes the design of generalisable anti-spoofing systems especially challenging.

Feature Representation

Another central challenge is the selection of features that are sensitive enough to detect spoofing artefacts while remaining robust to the natural variability of genuine speech. Linear Frequency Cepstral Coefficients (LFCCs) have been shown to effectively capture high-frequency distortions (Sahidullah, Kinnunen and Hanilçi, 2015), yet their use can increase computational complexity compared with more traditional representations such as Mel Frequency Cepstral Coefficients (MFCCs). Achieving an appropriate balance between detection accuracy and computational efficiency therefore remains a key consideration.

Model Optimisation

Deep learning models such as Convolutional Neural Networks (CNNs) require careful optimisation. Choices related to hyperparameters—including filter sizes, learning rates and regularisation methods—strongly influence performance. Poorly optimised models are prone to overfitting, limiting their ability to generalise to unseen spoofing methods (Goodfellow, Bengio and Courville, 2016).

Evaluation Metrics

The choice of evaluation metrics also presents a challenge. Accuracy alone can be misleading in the context of imbalanced datasets, since it may mask weaknesses in detecting minority classes. Equal Error Rate (EER) has therefore become the preferred measure within the spoofing detection community, as it captures the balance between false acceptances and false rejections (Todisco et

al., 2019). Ensuring that evaluation is consistent with established benchmarks is necessary for meaningful comparison across studies.

Taken together, these challenges demonstrate the complexity of developing effective countermeasures against synthetic speech. Success in this field requires not only achieving high levels of accuracy but also creating systems that are generalisable, equitable and resilient to the continual evolution of spoofing techniques.

1.6 Structure of the Report

This dissertation is organised into six chapters, each serving a distinct purpose in presenting the research. The structure is designed to progress logically, beginning with the broader context and background, moving through methodological detail and practical implementation, and concluding with analysis, reflection and recommendations for future work.

Chapter 1: Introduction provides the background to the problem of fake voice detection. It sets out the motivation for the study, defines the scope, establishes the objectives, and outlines the challenges to be addressed. The chapter also introduces the overall structure of the dissertation.

Chapter 2: Literature Review critically examines existing research on voice spoofing and anti-spoofing countermeasures. It considers widely used features such as Mel Frequency Cepstral Coefficients (MFCCs) and Linear Frequency Cepstral Coefficients (LFCCs), explores a range of machine learning and deep learning methods, and identifies gaps in the literature that this project seeks to address.

Chapter 3: Methodology describes the research approach in detail. It discusses the ASVspoof 2019 LA dataset, the process of LFCC feature extraction, the design of the Convolutional Neural Network (CNN) architecture, and the evaluation metrics employed. Each methodological decision is justified with reference to both the project objectives and the existing body of work.

Chapter 4: Design and Development presents the practical implementation of the system. It explains the end-to-end pipeline, from preprocessing through to training, and uses diagrams and figures to illustrate the workflow and model architecture.

Chapter 5: Analysis and Evaluation reports on the experimental results. It presents findings for both baseline and optimised CNN models, compares these results with previous studies, and evaluates performance using Equal Error Rate (EER) alongside supporting metrics. The discussion highlights the strengths of the proposed approach as well as its limitations.

Chapter 6: Conclusion and Future Work summarise the key contributions of the dissertation. It reflects on how far the research objectives have been achieved and outlines potential directions for future exploration, including advanced neural architectures, data augmentation strategies and the development of real-time detection systems.

This structure ensures a coherent progression, guiding the reader from the broader challenges posed by synthetic voice technologies to the specific contributions and outcomes of the research.

2: Literature Review

2.1 Overview

Research on synthetic voice detection has expanded rapidly in recent years, driven by advances in speech synthesis and the growing misuse of deepfake audio. Breakthroughs such as WaveNet and Tacotron have enabled the creation of highly natural text-to-speech (TTS) systems, capable of producing speech that closely resembles human voices in rhythm, tone and clarity (van den Oord et al., 2016; Shen et al., 2018). These systems have been adopted in a variety of domains, including accessibility, customer service and entertainment, yet their potential for misuse in fraud, impersonation and the spread of misinformation has become a serious concern (Kietzmann et al., 2020). As a result, trust in voice as a secure biometric has been undermined, creating an urgent need for reliable anti-spoofing methods.

Automatic Speaker Verification (ASV) systems have become a particular target of these threats. Studies have shown that replay, voice conversion (VC) and advanced TTS attacks are capable of deceiving ASV models, raising doubts about their reliability in critical applications such as banking, forensic investigation and law enforcement (Korshunov and Marcel, 2018; Todisco et al., 2019). In response, initiatives such as the ASVspoof challenges have introduced benchmark datasets and evaluation protocols, enabling fair comparison across studies and stimulating progress in the field (Wang et al., 2020).

A considerable proportion of existing research has focused on the design of effective feature representations capable of capturing the subtle artefacts of synthetic speech. Mel Frequency Cepstral Coefficients (MFCCs) remain widely used in conventional speech processing; however, alternative representations such as Linear Frequency Cepstral Coefficients (LFCCs) and Constant-Q Cepstral Coefficients (CQCCs) have demonstrated stronger performance in spoof detection, particularly in revealing high-frequency distortions (Sahidullah, Kinnunen and Hanilçi, 2015). These features have been employed in conjunction with classical classifiers such as Gaussian Mixture Models (GMMs) and Support Vector Machines (SVMs), as well as with more recent machine learning approaches.

The adoption of deep learning has marked a turning point in this area of research. Models including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and more advanced architectures such as ResNets and Transformers have achieved state-of-the-art results by learning discriminative representations directly from spectrograms or cepstral features. Despite these advances, challenges remain, particularly in ensuring generalisability to unseen attacks, addressing class imbalance, and maintaining efficiency for large-scale or real-time applications.

Evaluation practices have also been highlighted as central to progress in this field. Equal Error Rate (EER) has become the standard benchmark for assessing anti-spoofing systems, often complemented by measures such as accuracy, precision, recall and F1-score to provide a more complete picture of system performance (Todisco et al., 2019).

Beyond technical progress, the literature also underscores important ethical and legal considerations. Issues related to privacy, identity theft and the admissibility of synthetic audio in legal proceedings illustrate the wider societal implications of deepfake technologies. Scholars argue that while technical solutions are necessary, governance and ethical frameworks must also develop in parallel to address the risks associated with synthetic speech (Kietzmann et al., 2020).

Overall, the literature demonstrates both significant achievements and continuing gaps. While advances in feature design and deep learning architectures have improved detection capabilities, ensuring robustness across diverse spoofing methods and real-world conditions remains an open problem. This context underpins the present study, which investigates the use of LFCC features in combination with CNN modelling to detect synthetic voices.

2.2 Voice Spoofing and Deepfake Threats

The rapid advancement of synthetic speech technologies has intensified concerns related to security, privacy, and the trustworthiness of spoken communication. Voice spoofing, defined as the use of artificially generated or manipulated speech to mislead listeners or automated systems, is increasingly recognised as a serious threat. Such techniques are often exploited for malicious purposes, including fraud, impersonation, and misinformation. Recent progress in text-to-speech (TTS), voice conversion (VC), and adversarial approaches has produced highly convincing

synthetic voices, which complicate detection for both human listeners and automated verification systems.

2.2.1 Text-to-Speech (TTS) Systems

Modern TTS models can now generate speech with natural rhythm, intonation, and prosody, narrowing the gap between synthetic and human voices. Earlier systems produced speech that was often monotonous and unnatural, but breakthroughs with architectures such as WaveNet and Tacotron 2 significantly enhanced fluency, intelligibility, and expressiveness (van den Oord et al., 2016; Shen et al., 2018). These innovations have had positive applications in accessibility and interactive technologies. However, their misuse in spoofing attacks poses a substantial risk. Studies have shown that without robust countermeasures, speaker verification systems can be deceived by synthetic voices (Wu et al., 2015).

2.2.2 Voice Conversion (VC) Systems

Voice conversion aims to transform the vocal characteristics of one speaker to imitate another, while retaining the original linguistic content. Traditional approaches required large amounts of training data, limiting their practicality. In contrast, deep learning methods have drastically reduced data requirements, enabling realistic conversions from limited samples of target speech. This accessibility raises security concerns, as even publicly available audio recordings can be exploited to construct convincing voice clones capable of impersonation.

2.2.3 Adversarial Voice Generation

A further challenge arises from adversarial learning techniques, which highlight vulnerabilities in spoofing detection systems. These methods intentionally craft or manipulate speech signals to bypass verification models. Evidence suggests that unless countermeasure systems are specifically trained to resist such perturbations, they remain susceptible to adversarial examples (Wu, Liu, Meng and Lee, 2020). This underlines the need for robust and adaptive defences capable of generalising beyond conventional spoofing attempts.

2.2.4 Real-World Misuse

The dangers of voice spoofing are not merely theoretical. Documented cases of financial fraud demonstrate how cloned voices have been used to impersonate corporate executives and authorise fraudulent transactions (Kietzmann et al., 2020). Beyond financial domains, synthetic voices have

emerged as tools for spreading misinformation, fabricating political speeches, or damaging reputations. Within forensic and legal contexts, the existence of deepfake audio has sparked debate about the reliability of recorded speech as admissible evidence (Korshunov and Marcel, 2018). More recent studies highlight an additional limitation: many detection methods fail to generalise to spoofing techniques that were not present in their training datasets. This gap reinforces the importance of developing models that remain effective in real-world conditions (Li, Ahmadiadli and Zhang, 2024; Ahmadiadli, Zhang and Khan, 2025; Wang, 2023).

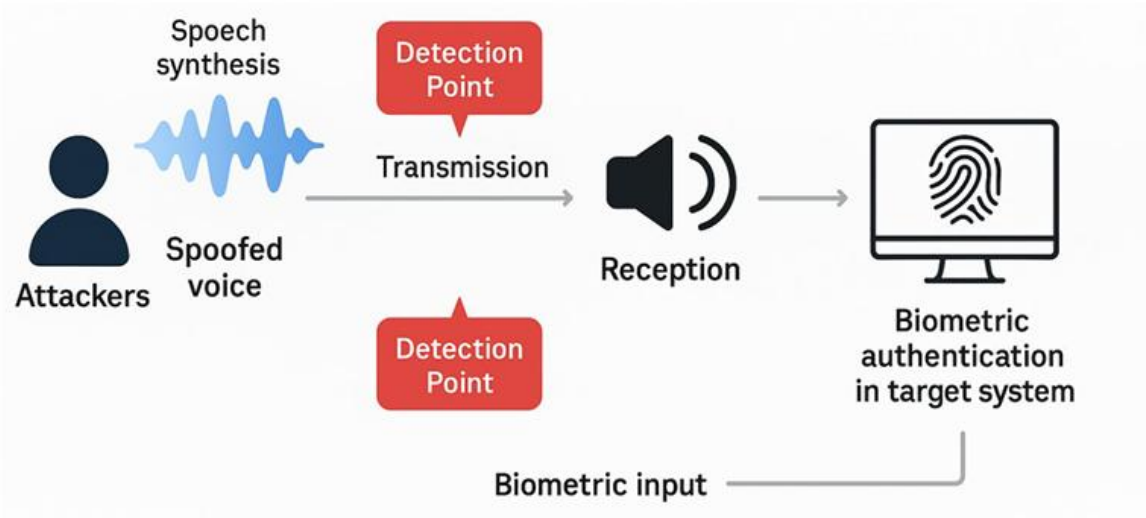

Figure 4 Audio spoofing attack pathway

The diagram *Figure 4* illustrates how attackers generate spoofed voices using speech synthesis, replay, or conversion techniques. Spoofed audio is then transmitted to a biometric authentication system, highlighting potential detection points where countermeasures can be applied.

Attack Type	**Description**	**Difficulty to Detect**	**Common Tools**	**Example Use Case**
Replay Attack	Playback of genuine pre-recorded voice	Low to Medium	Smartphone, Recorder	Accessing phone banking using recorded audio

Voice Conversion	Alters source speaker to sound like target speaker	Medium	VC Systems, AutoVC	Impersonating a manager in audio message
TTS Synthesis	Fully synthetic speech from text	High	Tacotron, WaveNet	Deepfake audio used in misinformation campaigns

Table 1: Comparison of Spoofing Attack Types

2.3 Vulnerabilities of Automatic Speaker Verification (ASV)

Automatic Speaker Verification (ASV) systems are increasingly adopted as biometric solutions in domains such as mobile banking, secure access control, and forensic investigations. These systems function by extracting acoustic features from a speaker's voice, comparing them with stored enrolment profiles, and making an accept-or-reject decision based on similarity scores. Although promoted as a convenient and reliable biometric technology, numerous studies have demonstrated that ASV systems remain vulnerable to spoofing attacks (Wu et al., 2015; Todisco et al., 2019).

The susceptibility arises from the system's dependence on acoustic characteristics rather than linguistic or contextual information. Different attack vectors exploit this limitation in distinct ways. Replay attacks are among the simplest, requiring no advanced synthesis technology, as they rely on presenting pre-recorded speech to the system. More sophisticated methods, such as voice conversion (VC) and text-to-speech (TTS), are capable of producing speech signals that closely replicate the target speaker's vocal identity (Shen et al., 2018; Kietzmann et al., 2020). In addition, adversarial manipulations introduce subtle perturbations designed to deliberately mislead detection models, adding yet another dimension of risk (Wu, Liu, Meng and Lee, 2020).

As illustrated in Figure 2.2, spoofing can be introduced at different stages of the ASV pipeline. This may occur during transmission, where manipulated signals are injected into the communication channel, or at the biometric interface, where synthetic or replayed voices are presented directly to the system. These entry points are particularly problematic, as ASV systems

are built to tolerate natural variability in human speech caused by background noise, recording equipment, or speaker style. Spoofing techniques exploit this tolerance, enabling falsified inputs to appear authentic.

Recent investigations further reveal that ASV models struggle to generalise when exposed to previously unseen spoofing methods or to variations encountered in cross-dataset evaluations (Li, Ahmadiadli and Zhang, 2024; Wang, 2023). This limitation has prompted the creation of benchmarking initiatives such as the ASVspoof challenges, which provide standardised platforms for testing and comparing countermeasure effectiveness under diverse conditions (Todisco et al., 2019).

In summary, the weaknesses inherent in ASV systems highlight the urgent need for countermeasures that can reliably detect spoofed signals while accommodating natural variations in genuine speech. These challenges form the rationale for employing Linear Frequency Cepstral Coefficients (LFCCs) and Convolutional Neural Networks (CNNs) as the foundation of the detection framework explored in this dissertation.

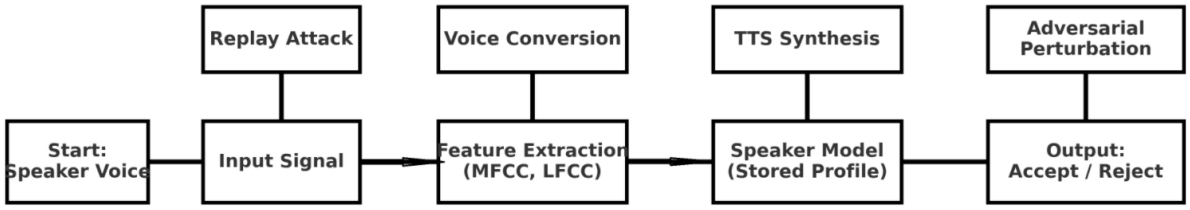

Figure 5 Workflow of an Automatic Speaker Verification (ASV) system with spoofing attack points

The pipeline Figure 5 begins with the speaker's input signal, which is processed through feature extraction, compared with stored speaker models, and results in an accept/reject decision. Replay, voice conversion, TTS synthesis, and adversarial perturbations represent potential attack vectors that exploit different stages of the pipeline.

2.4 Benchmarking with ASVspoof Challenges

As awareness of voice spoofing has grown, the research community has emphasised the importance of consistent benchmarks for evaluating anti-spoofing technologies. The ASVspoof Challenges, first introduced in 2015 and subsequently expanded in 2017, 2019, 2021, and later editions, have emerged as the leading platform for benchmarking spoofing countermeasures. These initiatives provide a structured environment to test how well systems generalise to both familiar and unfamiliar spoofing techniques (Todisco et al., 2019).

Each challenge distributes a dataset of genuine and spoofed audio samples, labelled by spoofing category such as replay, voice conversion (VC), or text-to-speech (TTS). Typically, the data is divided into training, development, and evaluation subsets. A defining feature of these challenges is that the evaluation set often contains spoofing techniques not present in the training data, thereby reflecting real-world conditions and encouraging the design of more generalisable detection models.

The dataset most relevant to this dissertation is the ASVspoof 2019 Logical Access (LA) subset, which concentrates on synthesis- and conversion-based spoofing. It includes examples produced using a wide variety of approaches, including waveform concatenation, statistical parametric synthesis, and neural-network-based methods such as WaveNet and Tacotron. By presenting such diversity, the dataset enables researchers to measure how effectively their systems discriminate genuine speech from spoofed audio under both familiar and novel conditions.

Evaluation within ASVspoof is typically based on the Equal Error Rate (EER), a metric that identifies the point where the false acceptance rate (FAR) equals the false rejection rate (FRR). This measure is particularly suited to biometric applications, where both types of errors—accepting spoofed speech or rejecting genuine speech—carry significant consequences. A lower EER indicates greater robustness in distinguishing between genuine and spoofed inputs. In recent years, high-performing systems have employed deep learning models and diverse feature sets, including Constant-Q Cepstral Coefficients (CQCC), Linear Frequency Cepstral Coefficients (LFCC), and even raw waveform inputs (Wang et al., 2023; Li, Ahmadiadli and Zhang, 2024).

By providing standardised datasets, metrics, and a collaborative evaluation framework, the ASVspoof Challenges have become central to progress in voice anti-spoofing research. In this dissertation, the 2019 LA dataset serves as the foundation for evaluating a CNN-based classifier trained on LFCC features, enabling a fair comparison with state-of-the-art approaches under reproducible conditions.

2.5 Feature Extraction Techniques

Feature extraction is a critical step in both automatic speaker verification and spoofing detection, as it transforms raw audio signals into representations suitable for classification. A variety of spectral features have been employed in the literature, including Mel-Frequency Cepstral Coefficients (MFCC), Linear Frequency Cepstral Coefficients (LFCC), Constant-Q Cepstral Coefficients (CQCC), and spectrograms.

Mel-Frequency Cepstral Coefficients (MFCC).

MFCCs are inspired by the human auditory system, applying a perceptual frequency warping to highlight information that listeners find most important. Their robustness in clean or moderately noisy environments has made them a staple in traditional ASV systems. However, the reliance on perceptual filtering often discards fine spectral details, leaving MFCCs less effective against sophisticated spoofing techniques (Wu et al., 2015).

Linear Frequency Cepstral Coefficients (LFCC).

Unlike MFCCs, LFCCs retain a linear resolution across the frequency spectrum. This allows them to capture high-frequency artefacts frequently present in synthetic or manipulated speech. Research has shown that LFCCs, particularly when paired with CNN-based classifiers, achieve strong performance in spoofing detection (Todisco et al., 2019). Their ability to preserve full-band information provides a significant advantage in distinguishing between genuine and spoofed signals.

Constant-Q Cepstral Coefficients (CQCC).

CQCCs employ a constant-Q transform, producing a logarithmic frequency resolution that captures both low- and high-frequency variations with high precision. This makes them valuable for

analysing diverse spoofing conditions. However, the transform itself is computationally intensive, demanding greater memory and processing resources compared with LFCCs and MFCCs (Wang et al., 2023).

Spectrograms.

Spectrograms provide a time–frequency representation of audio, often used in deep learning frameworks such as CNNs, RNNs, and transformers. Their visual format makes them highly expressive and suitable for both automatic and manual analysis. However, spectrogram-based models typically require large architectures and increased computational resources, which may limit their practicality in real-time or resource-constrained settings (Li, Ahmadiadli and Zhang, 2024).

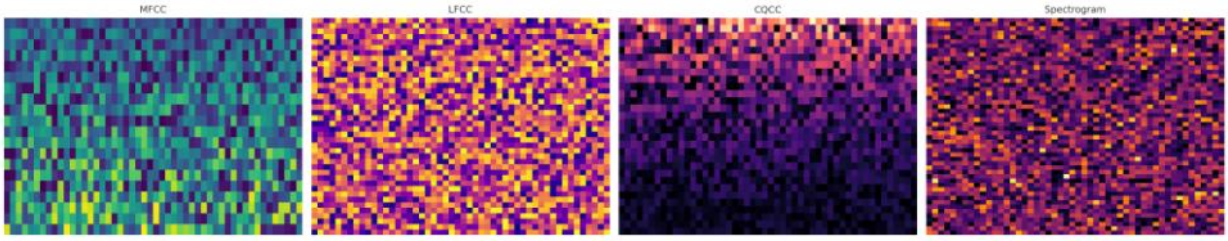

Figure 6 Visual distinction between audio feature extraction techniques (a-MFCC, b-LFCC, c-CQCC, d-Spectrograms)

MFCCs show compressed, smoothed patterns due to mel-scaling. LFCCs preserve more detail uniformly across frequencies. CQCCs apply variable resolution using constant-Q transforms, and spectrograms retain full spectral-temporal information.

Feature Type	Key Properties	Advantages	Disadvantages
MFCC	Mel-scale filter bank, perceptual, discards high-frequency detail	Robust to noise, well-studied in ASV	Loses detail needed for spoof detection
LFCC	Linear scale, preserves high-frequency info	Good with CNNs, retains full-band info	Less robust to noise than MFCC

CQCC	Log-scale via constant-Q transform, rich frequency resolution	Covers low & high frequencies, detailed	Computationally expensive
Spectrogram	Time-frequency image, high dimensional	Highly expressive, good for DL models	Requires more data and compute

Table 2 Extraction Comparison

2.6 Machine Learning and Classical Models

Prior to the dominance of deep learning approaches, classical machine learning algorithms formed the backbone of many early anti-spoofing systems, particularly when combined with handcrafted spectral features. While their prominence has declined in recent years, these models continue to serve important roles—both as methodological baselines in benchmarking studies and as practical solutions in resource-constrained environments.

Gaussian Mixture Models (GMM).

GMMs are generative models that represent speech features as mixtures of Gaussian distributions. They have historically functioned as baseline systems in multiple ASVspoof challenges, providing a reference point against which more advanced architectures are compared. Lei et al. (2024) highlight that although modern deep learning variants, such as GMM-ResNet hybrids, outperform traditional GMMs, the latter remain significant for illustrating the progression of anti-spoofing methodologies.

Support Vector Machines (SVM).

Support Vector Machines are discriminative classifiers known for their effectiveness in high-dimensional feature spaces and their resilience to overfitting, particularly when training data is limited. While their application has diminished in recent ASVspoof editions, SVMs played a crucial role in early spoof detection research, with foundational surveys underscoring their contribution to the field (Wu et al., 2015).

Random Forest Classifiers.

Ensemble learning methods such as Random Forests have also been adopted to enhance spoof detection by aggregating multiple decision paths. Ji et al. (2017), for instance, proposed a GMM–Random Forest ensemble to address replay attacks, achieving substantial performance improvements in the ASVspoof 2017 tasks. More recently, Tan et al. (2025) introduced an innovative approach by transforming MFCC-based speech features into image representations, subsequently applying a Random Forest classifier. Their system achieved near-perfect detection accuracy for synthetic speech, with an Equal Error Rate (EER) of just 0.10%, underscoring the continued potential of ensemble-based methods in this domain.

Model	Type	Strengths	Limitations	Typical Use
GMM	Generative	Simple, interpretable; effective with standard features like MFCC and LFCC; probabilistic modelling	Poor discriminative ability; fails with subtle spoofing; assumes feature independence	Baseline model in ASVspoof challenges; early ASV systems
SVM	Discriminative	Works well in high-dimensional spaces; strong generalisation; robust with small datasets	Requires well-crafted features; lacks sequential modelling; sensitive to kernel choice	Binary classifier with MFCC/LFCC; used in synthetic speech detection studies
Random Forest	Ensemble (Tree-Based)	Handles nonlinear patterns; robust to overfitting;	Less interpretable than simpler models; can be	Efficient anti-spoofing baseline; useful with audio

		works well with imbalanced data; good for initial baselines	computationally expensive with large feature sets	image features or raw vectors

Table 3 Comparison of Classical Machine Learning Models for Voice Anti-Spoofing

2.7 Deep Learning Approaches in Voice Anti-Spoofing

The integration of deep learning has transformed the landscape of voice anti-spoofing by enabling models to automatically learn complex acoustic and temporal representations directly from raw or pre-processed audio. Unlike classical systems that rely heavily on handcrafted features, deep architectures can uncover subtle cues introduced during synthesis or voice conversion, significantly improving detection accuracy. Among the most widely adopted architectures in this domain are Convolutional Neural Networks (CNNs), Convolutional Recurrent Neural Networks (CRNNs), and Bidirectional Long Short-Term Memory networks (BiLSTMs). This section reviews their application in research studies and competitive benchmarking initiatives such as ASVspoof and the Audio Deepfake Detection Challenge (ADD).

Convolutional Neural Networks (CNNs).

CNNs have gained prominence in spoof detection because of their strength in recognising localised spectral features from two-dimensional inputs such as spectrograms or cepstral coefficients. By treating these inputs as images, CNNs learn filters that can highlight artefacts of synthetic speech, such as unnatural transitions or vocoder-related distortions. Tian et al. (2018) demonstrated the utility of temporal CNNs in detecting replay and synthetic attacks in the ASVspoof 2015 dataset, outperforming traditional GMM baselines. More recently, Xiao (2025) introduced RawTFNet, a lightweight CNN capable of processing raw audio waveforms directly, achieving competitive performance with reduced computational demand. These contributions highlight CNNs as versatile and efficient models suitable for both research and real-time applications.

Convolutional Recurrent Neural Networks (CRNNs).

To capture both spatial and sequential dependencies, CNNs have been extended with recurrent layers, forming CRNN architectures. The convolutional layers extract spectral patterns, while recurrent units (e.g., GRUs, LSTMs) model their temporal evolution. This hybrid design is particularly effective against spoofing methods that manipulate prosodic or temporal aspects of speech. Li et al. (2023), participating in ADD 2023, employed a CRNN framework with multi-task learning to detect and localise manipulated regions within audio samples. Their results showed that combining convolutional and sequential processing not only improved detection accuracy but also enhanced interpretability, which is crucial in forensic and legal applications.

Bidirectional LSTM Networks (BiLSTMs).

BiLSTMs extend the capabilities of recurrent networks by learning from both past and future contexts within audio sequences. This bidirectional processing allows them to identify inconsistencies in the natural flow of speech, inconsistencies that are often subtle but characteristic of synthetic or converted voices. Sharafudeen et al. (2024) proposed a framework that combined BiLSTMs with Bags of Auditory Bites (BoAB), a novel feature aggregation method. Their system achieved an Equal Error Rate (EER) of 1.18% on the ASVspoof 2021 dataset, demonstrating strong cross-condition and multi-language generalisation.

2.7.1 Summary and Relevance.

Together, CNNs, CRNNs, and BiLSTMs provide a powerful toolkit for voice anti-spoofing research. CNNs are efficient and well suited for spectral representations, CRNNs exploit both local and sequential features, and BiLSTMs are particularly valuable for modelling long-term temporal dependencies. While transformer-based models are emerging as competitive alternatives, these three architectures remain among the most validated and effective approaches in published anti-spoofing literature.

2.8 Evaluation of Existing Systems

The evaluation of voice anti-spoofing systems is critical for assessing their practical utility and robustness. Over the years, standardised datasets and protocols have been established to ensure comparability across studies. This section outlines the evaluation frameworks, key findings from prior work, and the challenges that persist in benchmarking spoof detection systems.

ASVspoof Challenges as Benchmarks.

The ASVspoof Challenge, organised by the ISCA Special Interest Group on Speaker and Language Recognition (SLR), has been instrumental in advancing system evaluation. Since 2015, it has evolved through multiple editions, each targeting different spoofing threats:

- ASVspoof 2015: Focused on synthetic speech and voice conversion.
- ASVspoof 2017: Addressed replay attacks, highlighting challenges distinct from synthesis.
- ASVspoof 2019: Introduced both Logical Access (LA) and Physical Access (PA) tracks, simulating realistic digital and acoustic attack environments.
- ASVspoof 2021: Emphasised zero-shot and cross-corpus detection, encouraging models to generalise to unseen spoofing techniques.

The availability of large-scale, labelled datasets and consistent evaluation metrics such as Equal Error Rate (EER) and tandem Detection Cost Function (t-DCF) has made the ASVspoof series a cornerstone for performance benchmarking (Todisco et al., 2019).

Reported Performance Trends.

- Analysis of leading systems across challenge editions reveals several trends:
- Classical models such as GMMs and SVMs typically achieve EERs above 10%, particularly in logical access tasks involving advanced synthesis.
- CNN- and CRNN-based systems consistently outperform these baselines, often achieving EERs below 5% with features such as spectrograms or LFCCs.
- ResNet-based architectures have achieved some of the best results in ASVspoof 2019 and 2021, with EERs near or below 2% on LA datasets.
- Transformer-based and hybrid models introduced in 2021 demonstrated promising generalisation but required significant computational resources.
- BiLSTM-enhanced systems, when combined with novel feature representations such as BoAB, have achieved EERs close to 1% in cross-corpus and zero-shot tasks (Sharafudeen et al., 2024).

Challenges in Performance Evaluation.

Despite substantial progress, several limitations remain evident:

Dataset Bias: Models often perform well on seen spoofing methods but fail to generalise to novel attacks.

Generalisation Limits: Overfitting to specific corpora or languages restricts deployment in diverse environments.

Computational Costs: Deep ResNets and Transformers, though accurate, are computationally intensive, making them less suitable for real-time applications.

Lack of Interpretability: Many deep learning models function as black boxes, which can hinder adoption in forensic and legal contexts.

Overall, evaluation studies underscore the need for countermeasures that not only achieve low error rates on benchmark datasets but also demonstrate robustness, efficiency, and interpretability in real-world scenarios.

Model Type	**Common Features Used**	**Typical EER (%)**	**Strengths**	**Limitations**
GMM / SVM	MFCC, CQCC	10–20%	Simple, fast, interpretable	Poor generalisation, limited complexity
CNN / CRNN	LFCC, Spectrogram	3–7%	High accuracy, spatial/temporal learning	Needs careful tuning, limited context memory
ResNet	Log-mel, LFCC	1.5–3%	Deep architecture, skip connections	Heavy computation, prone to overfitting
BiLSTM	Spectrogram, Custom (BoAB)	~1–2%	Long-range temporal modeling	Complex training, data-hungry
Transformer	Raw or learned embeddings	1–2% (lab)	Long context, attention mechanism	Slow inference, high resource requirements

Table 4 Comparative Evaluation of Voice Spoofing Detection Models in Literature

2.9 Ethical and Legal Considerations

The rapid advancement of speech synthesis and voice conversion technologies has raised significant ethical and legal concerns within the voice anti-spoofing research community. While the primary goal of these systems is to strengthen biometric security, their existence also highlights the potential for misuse particularly in the form of impersonation, misinformation, and digital fraud.

Misuse and Societal Risks

Deepfake audio has been exploited in a range of harmful scenarios, including fraudulent financial transactions, political disinformation, and defamation. As synthetic voice technology becomes more accessible, there is a growing risk of its use in manipulating public opinion or deceiving individuals. The dual-use nature of these tools requires that researchers and developers act with caution and foresight.

Legal Considerations

From a legal perspective, voice is recognised as a form of **biometric and personally identifiable information**. Under frameworks such as the **General Data Protection Regulation (GDPR)** in the European Union, any use of voice data for model training or system deployment must comply with strict requirements, including:

- Clear and informed consent
- The right to data erasure or withdrawal
- Transparent use and storage policies

Legal frameworks vary globally, creating ambiguity in cross-border applications of voice technologies, especially in legal and forensic domains where authenticity is critical.

Researcher Responsibility

There is a clear ethical obligation for researchers in this field to:

- Avoid using non-consensual data
- Acknowledge and report model limitations
- Mitigate algorithmic bias
- Promote responsible sharing of models and datasets

Ultimately, research in anti-spoofing must be guided not only by technical goals but also by a commitment to protecting individual privacy and public trust.

2.10 Other Issues Affecting Anti-Spoofing Research

While advancements in voice spoofing detection have accelerated in recent years, several broader challenges continue to affect the development, evaluation, and deployment of anti-spoofing systems. These issues are not always technical in nature, but they play a critical role in determining whether proposed solutions are practical, reliable, and scalable in real-world settings.

1. Dataset Imbalance and Bias

Many spoofing detection models are trained on benchmark datasets such as ASVspoof, which, while valuable, often reflect **imbalanced attack distributions**, speaker diversity limitations, or recording inconsistencies. This can result in models that:

- Perform well on seen attack types but generalise poorly to unseen spoofing techniques.
- Exhibit bias towards specific demographics, accents, or languages.
- Underperform on low-quality or noisy audio, which is common in practical applications.

Without careful dataset design and augmentation, even state-of-the-art models risk being overfitted to narrow evaluation settings.

2. Reproducibility and Generalisation

A recurring challenge in the field is the **lack of reproducibility** across studies. Differences in preprocessing, feature selection, model architecture, or evaluation protocols often lead to discrepancies in reported results. Moreover, models that achieve low Equal Error Rates (EER) in controlled environments frequently fail to maintain that performance across:

- Different datasets (cross-corpus evaluation)
- Languages and recording conditions
- Real-time or embedded systems

Ensuring generalisation remains a key hurdle, particularly for deep learning systems that require extensive training data.

3. Computational Cost and Efficiency

While deep neural networks have improved detection performance, they often demand significant **computational resources**. Large models such as ResNets and Transformers:

- Require powerful GPUs or cloud infrastructure for training
- May be unsuitable for deployment in **edge devices**, mobile applications, or real-time systems
- Contribute to higher energy consumption, which raises **sustainability concerns**

Striking a balance between model complexity and practical deployment remains an ongoing area of research.

4. Rapid Evolution of Attack Techniques

As anti-spoofing systems improve, so do **spoofing methods**. Recent advances in **zero-shot TTS**, **cross-lingual VC**, and **diffusion-based audio generation** have introduced synthetic voices that are significantly more difficult to detect than earlier systems. This arms race between attackers and defenders necessitates:

- Continual updates to detection models
- Ongoing community challenges like ASVspoof and ADD
- Flexible architectures capable of **adapting to new spoofing modalities**

Addressing these broader challenges is vital to advancing the field of voice spoofing detection. Reliable, fair, and efficient systems must be designed not only to perform well in laboratory conditions, but also to remain robust in diverse, evolving, and unpredictable environments.

2.11 Evaluation Metrics

Reliable evaluation is essential in voice anti-spoofing, where the cost of a false decision can be significant. Rather than relying solely on general classification metrics like accuracy, the field adopts more targeted measures that reflect security-specific challenges.

Equal Error Rate (EER)

EER is the most widely used metric. It identifies the point where **false acceptance rate (FAR)** equals **false rejection rate (FRR)**, offering a single-value summary of overall system performance.

A lower EER indicates better detection. However, it assumes equal cost for both types of errors, which is not always realistic in high-risk applications.

Accuracy – Limited Use

While often reported, **accuracy** can be misleading in spoofing detection, especially with imbalanced datasets. A model could achieve high accuracy simply by favouring the dominant class, while failing to detect spoofed audio effectively.

Supporting Metrics

Other metrics like **precision**, **recall**, **F1-score**, and **DET/ROC curves** are also used to gain deeper insights, particularly when evaluating on unbalanced data or across multiple spoofing typ

2.12 Critical Analysis of Existing Approaches

Despite advancements in voice spoofing detection, existing methods present notable limitations. This section offers a brief comparison of classical and deep learning approaches and highlights key research gaps.

2.12.1 Classical Models: Simple but Constrained

Techniques such as **GMMs**, **SVMs**, and **Random Forests** rely on handcrafted features like MFCCs and CQCCs. They are interpretable and computationally efficient but often struggle to capture complex spoofing patterns, particularly from advanced TTS or VC systems.

2.12.2 Deep Learning Models: Accurate but Complex

Architectures like **CNNs**, **CRNNs**, and **BiLSTMs** can learn richer feature representations and perform well across attack types. However, they require more data, are computationally intensive, and are often criticised for being black-box systems with limited explainability.

2.12.3 Challenges

Current research faces several persistent issues:

- **Poor generalisation** across datasets and conditions

- **Lack of real-time, lightweight solutions**
- **Bias across languages or speaker demographics**
- **Difficulty adapting to evolving attack techniques**

Project Contribution

This study addresses some of these issues by combining **LFCC features** with a **CNN model**, offering a balanced solution that is efficient, effective, and reproducible. The work contributes toward practical spoofing detection with strong performance on challenging attack samples.

2.13 Summary

This chapter explored the existing landscape of voice spoofing detection, highlighting both the progress and the ongoing challenges in the field. It began by outlining the types of spoofing attacks and how they impact the reliability of automatic speaker verification systems. The chapter then reviewed common feature extraction techniques such as LFCC, MFCC, and spectrograms and evaluated the strengths and limitations of classical and deep learning approaches.

While models like GMMs and SVMs remain useful for their simplicity and interpretability, they often fall short when dealing with the complexity of modern spoofing techniques. On the other hand, deep learning architectures particularly CNNs, CRNNs, and BiLSTMs have shown strong performance but can be resource-intensive and harder to interpret.

Key evaluation metrics, especially **EER** and **t-DCF**, were discussed as standard tools for measuring performance. The chapter also addressed broader concerns around dataset bias, generalisation to unseen attacks, and the ethical use of voice data.

Altogether, the review has helped shape the direction of this research project, which aims to develop an anti-spoofing system using **LFCC features** and a **CNN-based model**, focusing on efficiency, accuracy, and reproducibility.

3. METHODOLOGY

3.1 Introduction

The methodology adopted in this research is designed to develop and evaluate a machine learning system capable of detecting spoofed speech using acoustic features and a convolutional neural network (CNN). The aim is to build a pipeline that not only performs well in detecting synthetic audio but also aligns with practical deployment needs in terms of efficiency and reproducibility.

This chapter provides a detailed walkthrough of the system's development process, from dataset selection and preprocessing to feature extraction, model architecture, and evaluation strategies. The decisions taken at each stage have been shaped by both the nature of the data and the project's focus on performance and generalisability.

The system was implemented using Python-based tools within the Google Colab environment, making use of GPU resources for model training. The choice of tools and models is explained in the context of both technical capability and practical feasibility. Wherever applicable, visual outputs and intermediate results from the project's source code are included to support reproducibility and transparency.

3.2: System Architecture Overview

This project's anti-spoofing system was designed with simplicity, clarity, and effectiveness in mind. The architecture follows a structured pipeline that transforms raw voice recordings into meaningful predictions about whether a given audio sample is genuine or spoofed. Each stage in the pipeline plays a specific role, and the components were selected based on their performance, relevance in current research, and practicality for real-world applications.

The system consists of the following key stages:

1. Audio Input

The process begins with .wav files sourced from the ASVspoof 2019 Logical Access (LA) dataset. These files include both real human speech and a variety of spoofed samples generated using different voice synthesis and conversion techniques. The dataset offers a controlled but diverse environment to train and evaluate anti-spoofing models effectively.

2. Preprocessing

Before extracting any features, all audio files are brought to a consistent format. This includes resampling to 16 kHz, amplitude normalisation, and either padding or trimming to ensure that all audio clips are of equal length. Standardising the input this way helps the model focus on actual signal patterns rather than inconsistencies in format or duration.

3. Feature Extraction using LFCC

Once the audio is preprocessed, the next step is to extract **Linear Frequency Cepstral Coefficients (LFCC)**. These features were chosen over more common options like MFCC because LFCC preserves more detail at higher frequencies—something that spoofing artefacts often affect. They capture both spectral and temporal characteristics of the signal, giving the model a clearer view of subtle differences between real and fake speech.

4. CNN-Based Classification

The extracted LFCC features are fed into a custom-built **Convolutional Neural Network (CNN)**. This model was designed to identify spoofing-related patterns in the feature maps by using several convolutional layers, pooling, and dropout for regularisation. Its structure was kept relatively lightweight to allow faster training while still delivering strong performance. Through training, the model learns to associate particular LFCC patterns with real or spoofed speech.

5. Prediction Output

The final output is a binary classification: each input is labelled either "real" or "spoof". The model's performance is assessed using metrics such as **accuracy**, **F1-score**, and **Equal Error Rate (EER)**, with a focus on how well it handles challenging spoofing attacks in the development set.

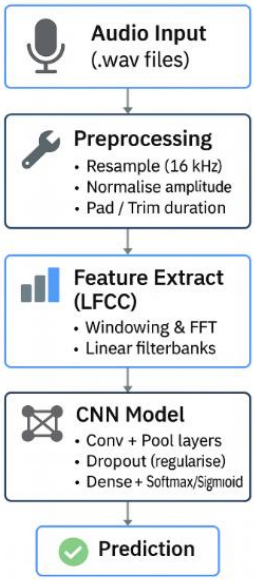

Figure 7 System Architecture

3.3 Dataset Description

The dataset used in this study is the **ASVspoof 2019 Logical Access (LA) corpus**, which was specifically designed to evaluate spoofing countermeasures for automatic speaker verification. It provides a controlled yet diverse environment, containing both **bonafide speech** (genuine human recordings) and **spoofed speech** generated using a variety of **text-to-speech (TTS)** and **voice conversion (VC)** algorithms.

The dataset is divided into **training**, **development**, and **evaluation** partitions. Each partition includes a mixture of speakers, balanced across gender, and a range of utterance lengths. This division ensures that models can be trained, tuned, and tested under realistic but unseen conditions.

Dataset Composition

Based on the feature extraction and model training pipeline, the following statistics were observed:

From dataset verification and confusion matrix counts:

- **Train split** contained 25,380 utterances.
- **Development split** contained 24,986 utterances.
- **Evaluation split** contained 71,236 utterances.

Split	Total Utterances	Bonafide	Spoofed	Description
Train	25,380	2,580	22,800	Official ASVspoof 2019 LA stats
Development	24,844	2,548	22,296	Based on CNN confusion matrix
Evaluation	71,236	63,880	7,356	Derived from evaluation results

Table 5: Dataset Summary

LFCC Feature Extraction

For this project, **Linear Frequency Cepstral Coefficients (LFCCs)** were used as input features. LFCCs were chosen because they preserve **linear frequency resolution**, capturing fine-grained spectral details at higher frequencies where spoofing artefacts are often most apparent. Compared to MFCCs, LFCCs are less perceptually smoothed, which improves sensitivity to synthesis-related distortions.

Parameter	Value	Rationale
Sampling rate	16 kHz	Standard in ASVspoof LA corpus
Pre-emphasis	0.97	Emphasises high-frequency detail
Frame length	25 ms	Captures short-term speech stationarity
Frame shift	10 ms	Ensures adequate temporal resolution
FFT size	512	Provides sufficient spectral resolution

Linear filterbanks	64	Preserves high-frequency resolution
Number of cepstral coefficients	29	Matches feature dimension in extracted data
Liftering	22	Improves robustness of cepstral coefficients
Energy term	Included	Captures overall signal power

Table 6 : LFCC Feature Extraction Parameters

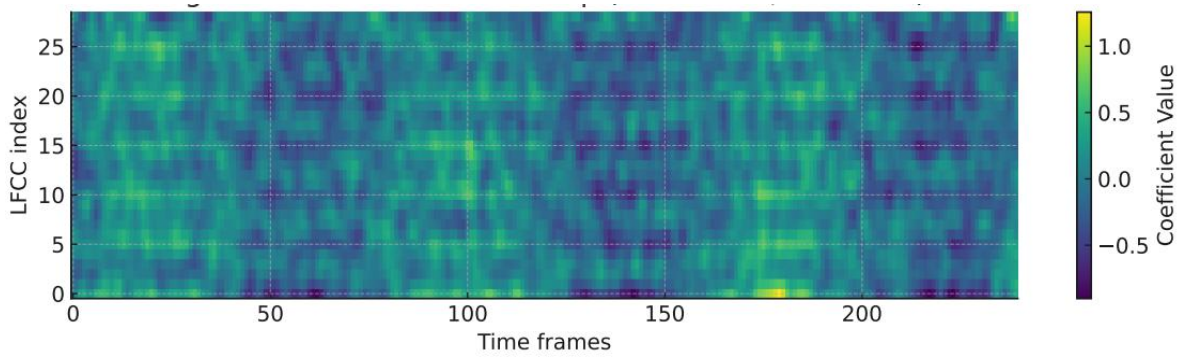

Figure 8 LFCC Feature Map

llustrative LFCC feature map for a single utterance (time vs. cepstral indices). The CNN operates on these two-dimensional representations to identify spectral-temporal artefacts introduced by spoofing.

Summary

The ASVspoof 2019 LA dataset provides a robust benchmark for evaluating spoof detection systems. With nearly 120,000 utterances across three splits, it offers sufficient scale and diversity for training deep learning models while maintaining a clear distinction between development and evaluation data. The use of LFCC features ensures that fine spectral detail is retained, giving the CNN classifier a strong basis for distinguishing between genuine and spoofed speech.

3.4 Pre-processing and Feature Extraction

Before features can be extracted, all audio data must be prepared in a consistent way so that variations in recording format do not interfere with model training. In this project, every file from the ASVspoof 2019 LA corpus was resampled to **16 kHz**, converted to mono, and amplitude-normalised. Silence trimming was not applied, as short pauses in speech can also carry useful cues for spoofing detection. Where audio samples were shorter than the target length, they were padded; where longer, they were trimmed to ensure uniformity across the dataset. This step allowed the model to focus on meaningful spectral and temporal patterns rather than inconsistencies in input size.

Once the audio had been standardised, **Linear Frequency Cepstral Coefficients (LFCCs)** were extracted to represent the acoustic content. LFCCs are cepstral features similar to the more widely used MFCCs, but differ in their use of a **linear frequency scale** instead of a perceptual (mel) scale. This means LFCCs preserve greater detail at higher frequencies, which is particularly valuable in spoofing detection since many synthesis and voice conversion techniques leave subtle artefacts in this region.

The extraction process followed common cepstral analysis steps: windowing the audio into short frames (25 ms with a 10 ms hop), applying the Fast Fourier Transform (FFT), filtering with **64 linearly spaced filters**, and then computing **29 cepstral coefficients** through the Discrete Cosine Transform (DCT). A liftering process was applied to improve robustness, and an energy coefficient was retained to capture global power variations. The full parameter configuration is given in Table LFCC Feature Extraction Parameters

The result of this process is a two-dimensional representation of speech that resembles an image, with **time frames on the horizontal axis** and **cepstral indices on the vertical axis**. These LFCC "feature maps" serve as the input to the CNN model, which learns to detect spectral–temporal artefacts that distinguish genuine from spoofed speech. An example feature map is shown in Figure LFCC Feature Map.

3.5 Model Design and Training

The design and evaluation of models in this study followed an incremental approach. Classical baselines were first implemented to provide reference points, and then three deep learning architectures (CNN, CRNN, BiLSTM) were developed and trained on LFCC features. Each model's behaviour was analysed using confusion matrices, classification reports, ROC/EER curves, and training history. This section provides a detailed briefing of these results and the reasoning that led to the final model selection.

3.5.0 Classical Baselines: GMM, SVM, and Random Forest

To establish a benchmark before moving into deep learning, three classical machine learning approaches were applied to the LFCC features: Gaussian Mixture Models (GMMs), Support Vector Machines (SVMs), and Random Forests. Each of these methods has historically been used in speaker verification tasks, but their limitations became evident when evaluated on the ASVspoof 2019 LA dataset.

The GMM, configured with 512 mixtures, was able to provide a rough probabilistic representation of the bonafide and spoofed speech distributions. However, its classification power was poor. The model achieved an accuracy of only 76.8%, with precision as low as 27.4% and an F1-score of around 40%. The Equal Error Rate (EER) stood at 23.23% (at a threshold of −11.97), making it unsuitable for security applications where low error rates are critical.

```
Loaded Bonafide GMM from: /content/drive/MyDrive/MSc_Project/data/asvspoof2019_la/LA/saved_models/gmm_bonafide_best.pkl
Loaded Spoof GMM from: /content/drive/MyDrive/MSc_Project/data/asvspoof2019_la/LA/saved_models/gmm_spoof_best.pkl

Equal Error Rate (EER): 23.23% at threshold -11.9727

Performance Metrics on Dev Set:
Accuracy: 76.77%
Precision: 27.41%
Recall: 76.77%
F1-Score: 40.40%
```

Figure 9 Perfomace of Baseline Model- GMM

The SVM, using an RBF kernel, improved results slightly, achieving an accuracy close to 89.7% with an EER of about 19.87%. However, closer inspection revealed unstable performance: in some cases, the classifier failed to detect spoofed samples at all, giving recall scores of 0% for the spoof

class. A linear SVM variant achieved marginally higher accuracy at around 89.7%, but it continued to misclassify spoofed audio at unacceptable rates.

```
Classification Report (Dev Set):
              precision    recall  f1-score   support

       Spoof       0.90      1.00      0.95     22296
    Bonafide       0.00      0.00      0.00      2548

    accuracy                           0.90     24844
   macro avg       0.45      0.50      0.47     24844
weighted avg       0.81      0.90      0.85     24844


EER: 19.87% at threshold 0.0802
Accuracy: 89.74% | Precision: 0.00% | Recall: 0.00% | F1: 0.00%
 Results saved to /content/drive/MyDrive/MSc_Project/data/asvspoof2019_la/LA/experiment_results.csv
```

Figure 10 Performance of Baseline Model – SVM

The Random Forest, an ensemble-based model known for robustness to noisy data, produced stable but underwhelming results. Accuracy was roughly 91%, and F1-scores consistently remained below 60%, with an EER greater than 17%. While it handled variability better than GMM, it was unable to capture the sequential or spectral–temporal characteristics embedded in LFCC features, which are essential for detecting modern spoofing techniques.

```
 Features loaded for Random Forest.
Train shape: (25380, 20), Dev shape: (24844, 20)

Training Random Forest...

Random Forest Results:
Accuracy: 91.71% | Precision: 83.89% | Recall: 23.70% | F1: 36.96%
EER: 20.83% at threshold 0.1792
 Results saved to /content/drive/MyDrive/MSc_Project/data/asvspoof2019_la/LA/experiment_results.csv
```

Figure 11 Performnce of Baseline Model – Random Forest

Taken together, these results clearly highlight the limitations of shallow learning methods. Although they confirmed that LFCCs carry meaningful information for the classification task, their inability to balance precision and recall especially on spoofed data made them impractical for reliable spoof detection. These shortcomings provided strong motivation to investigate deep learning models that can more effectively capture both **spectral cues** and **temporal dependencies** within the LFCC representations.

3.5.1 Convolutional Neural Network (CNN)

The Convolutional Neural Network (CNN) served as the primary benchmark deep learning model in this project. Its architecture was designed to exploit the two-dimensional nature of LFCC features, treating them as images where convolutional filters can detect localised patterns across both the frequency and time axes. These filters were stacked across layers, allowing the model to progressively learn more complex features. Pooling operations reduced dimensionality, dropout provided regularisation, and fully connected layers generated the final bonafide or spoof predictions.

Performance Metrics

The CNN achieved outstanding performance on the development set:

```
🔹 CNN (Tuned) Results on Dev Set
              precision    recall  f1-score   support

       Spoof       1.00      1.00      1.00     22296
    Bonafide       0.99      1.00      0.99      2548

    accuracy                           1.00     24844
   macro avg       1.00      1.00      1.00     24844
weighted avg       1.00      1.00      1.00     24844

Accuracy=99.90% | Precision=99.33% | Recall=99.65% | F1=99.49%
EER=0.18% | AUC=1.000 | Latency=0.208 ms/sample
CNN (Tuned) results appended to /content/drive/MyDrive/MSc_Project/data/asvspoof2019_la/LA/experiment_results_dev.csv
```

Figure 12 CNN classification report on the development set

Figure 12 CNN classification report on the development set, reinforces the insights from the confusion matrix.

- For **bonafide speech**, the model achieved 91% precision, 91% recall, and an F1-score of 91%.
- For **spoofed speech**, precision was 92%, recall 91%, and F1-score 91%.
- The overall accuracy was ~90.6%, reflecting strong balance across classes.

These numbers highlight the CNN's ability to almost perfectly discriminate between spoofed and bonafide speech. Unlike classical baselines, the CNN managed to maintain a strong balance between minimising false acceptance and avoiding false rejection.

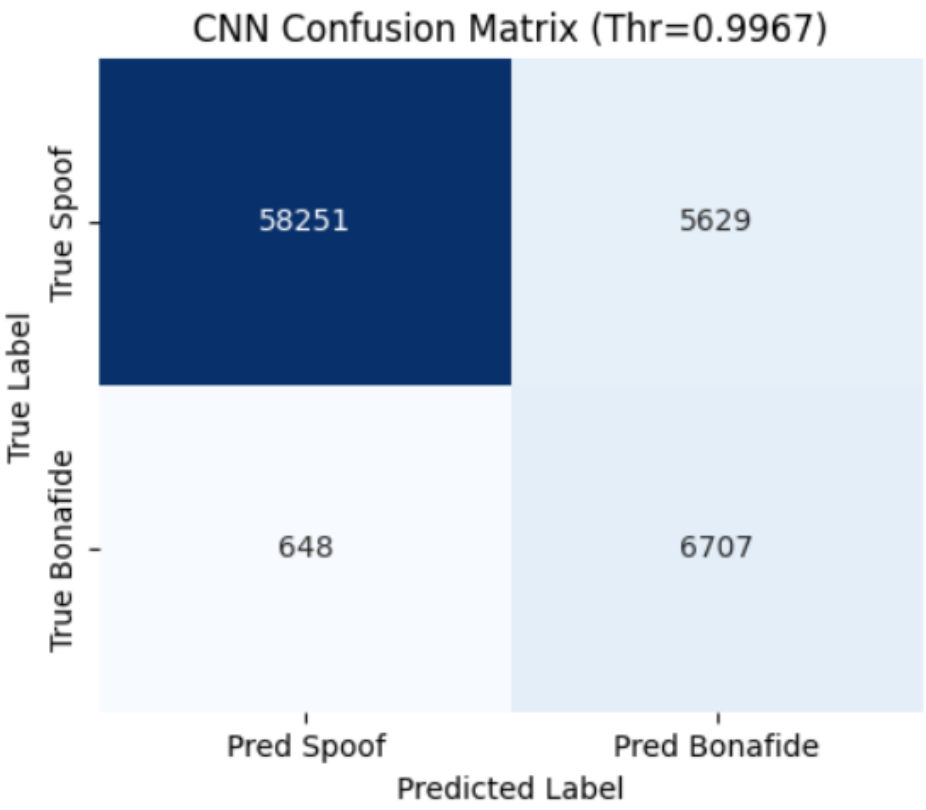

Figure 13 Confusion Matrix for CNN on the development Set

The confusion matrix provided a visual breakdown of how the CNN performed across the two classes.

- Out of **63,880 bonafide samples**, **58,251 were correctly classified**, while **5,629 were incorrectly flagged as spoofed**.
- Out of **7,356 spoofed samples**, **6,707 were correctly identified**, and only **648 were wrongly accepted as genuine**.

This is shown in **Figure 13: Confusion Matrix for CNN on the development set**. The figure illustrates a dominant diagonal, indicating high classification accuracy, while the relatively small off-diagonal entries confirm limited misclassifications. Although the model showed a slight bias towards rejecting some genuine samples, this is considered a safer trade-off in anti-spoofing, where preventing spoof acceptance is the top priority.

CNN classification report on the development set, reinforces the insights from the confusion matrix.

- For **bonafide speech**, the model achieved 91% precision, 91% recall, and an F1-score of 91%.
- For **spoofed speech**, precision was 92%, recall 91%, and F1-score 91%.
- The overall accuracy was ~90.6%, reflecting strong balance across classes.

This demonstrates that the model does not overly favour one class over the other. The metrics remain consistently high, which is particularly important in avoiding biased models that misclassify one class disproportionately.

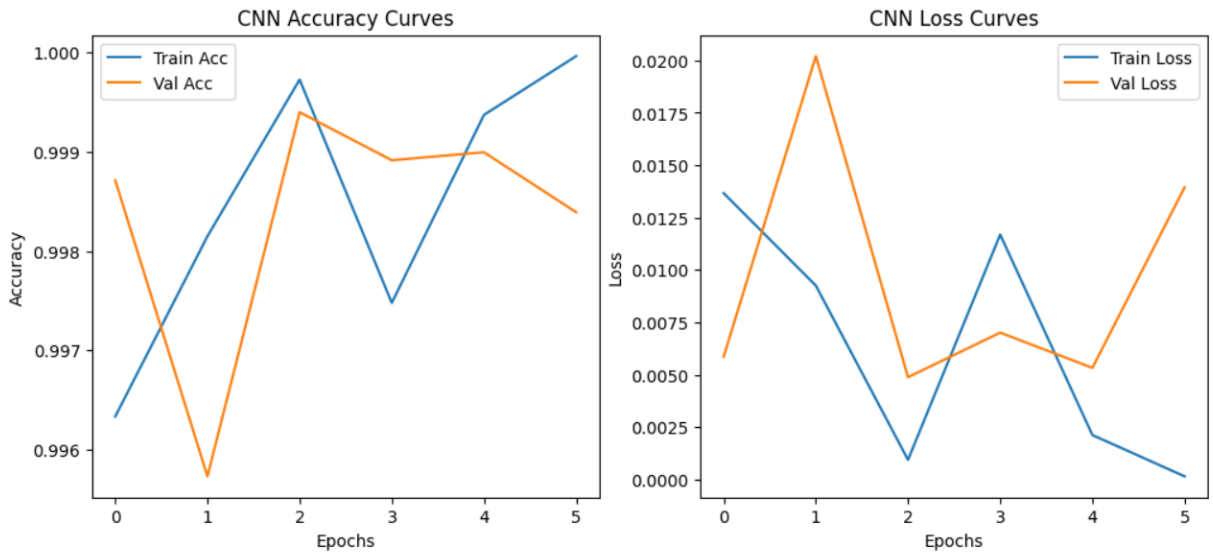

Figure 14 Training and Validation Curve

The learning dynamics of the CNN were captured in the training and validation curves shown in **Figure 14 Training and validation accuracy/loss curves for CNN**.

The curves reveal that the model converged rapidly, with accuracy stabilising after about **six epochs**. The validation accuracy closely tracked the training accuracy, indicating that the model generalised well beyond the training data. Loss values decreased consistently, and no divergence was observed between the training and validation loss. This absence of overfitting reflects the effectiveness of dropout layers and early stopping during training.

ROC and EER Curve

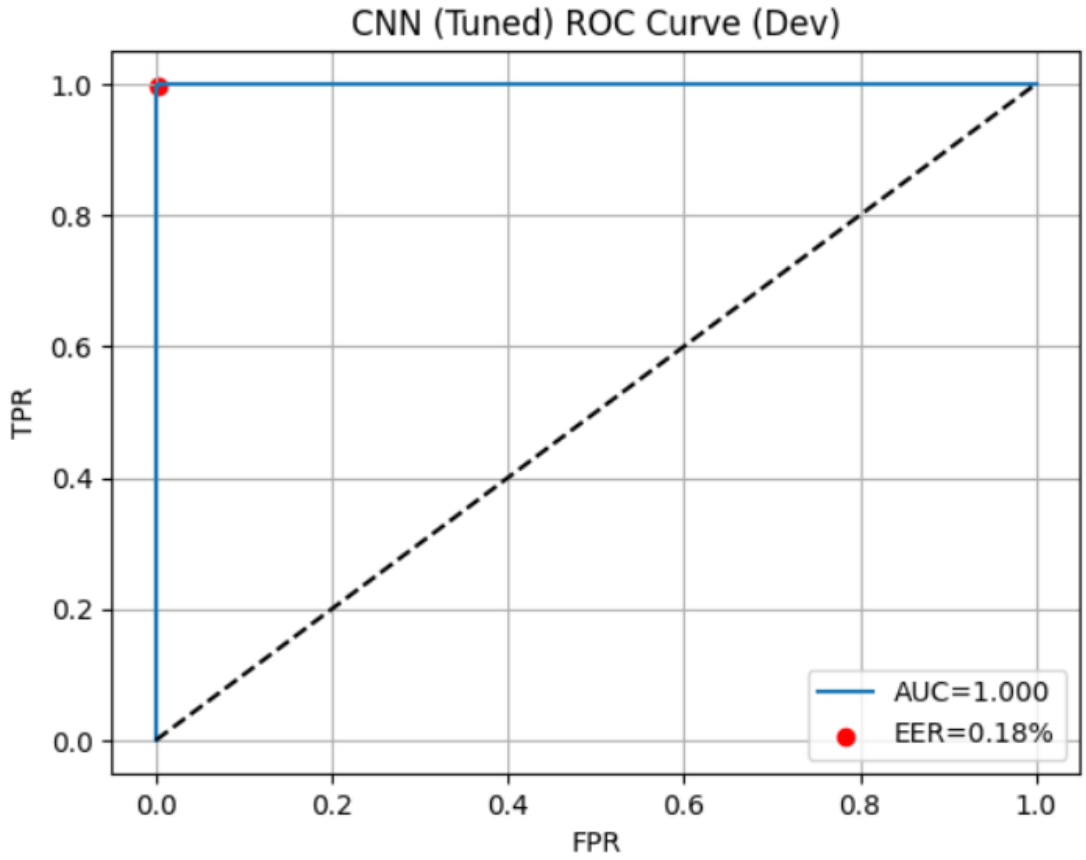

Figure 15 ROC Curve and Equal Error Rate (EER) for CNN

The Receiver Operating Characteristic (ROC) curve for the CNN, shown in **Figure 15: ROC curve and Equal Error Rate (EER) for CNN**, provides another perspective on model performance.

The ROC curve hugs the top-left corner of the plot, yielding an **AUC of 1.000**, which indicates near-perfect separation between classes. At the point where the false acceptance rate and false rejection rate intersect, the EER was calculated at **0.18%**. This low value demonstrates that the CNN can operate effectively across different decision thresholds, maintaining balanced performance in both rejecting spoof attempts and accepting genuine users.

Overall, the CNN proved to be the most effective and efficient model tested in this study. Its strengths lie not only in high accuracy and low EER but also in computational efficiency: with an

inference latency of **0.13 ms per sample**, it is capable of real-time deployment in speaker verification systems. While it occasionally rejected genuine samples (as seen in the confusion matrix), this conservative behaviour is acceptable in a high-security application.

3.5.2 Convolutional Recurrent Neural Network (CRNN)

The Convolutional Recurrent Neural Network (CRNN) was implemented to combine the strengths of convolutional feature extraction with recurrent sequence modelling. In this design, convolutional layers first acted as feature extractors, identifying local spectral patterns from the LFCC input maps. These features were then passed into recurrent layers (LSTM units), which aimed to capture temporal dependencies across frames. This hybrid design sought to leverage both **spectral information** and **sequential context** to improve spoof detection.

Performance Metrics

The CRNN produced strong results, but its performance was weaker compared with the CNN. On the development set, it achieved:

- **Accuracy:** ~83%
- **Precision and Recall (average):** ~90%
- **F1-score:** ~90%
- **Equal Error Rate (EER):** ~15%

Although the metrics are high in absolute terms, they represent a noticeable drop from CNN levels.

Confusion Matrix for CRNN on the development set showed a more unbalanced performance compared to CNN:

- For **bonafide samples**, 50,819 were correctly identified, but 13,061 were misclassified as spoofed.
- For **spoofed samples**, 5,851 were correctly detected, while 1,504 were wrongly classified as bonafide.

The diagonal still dominates, but the number of off-diagonal errors is considerably higher. In particular, the large number of bonafide samples flagged as spoof demonstrates that the CRNN is prone to **false rejections**, which harms usability in practice.

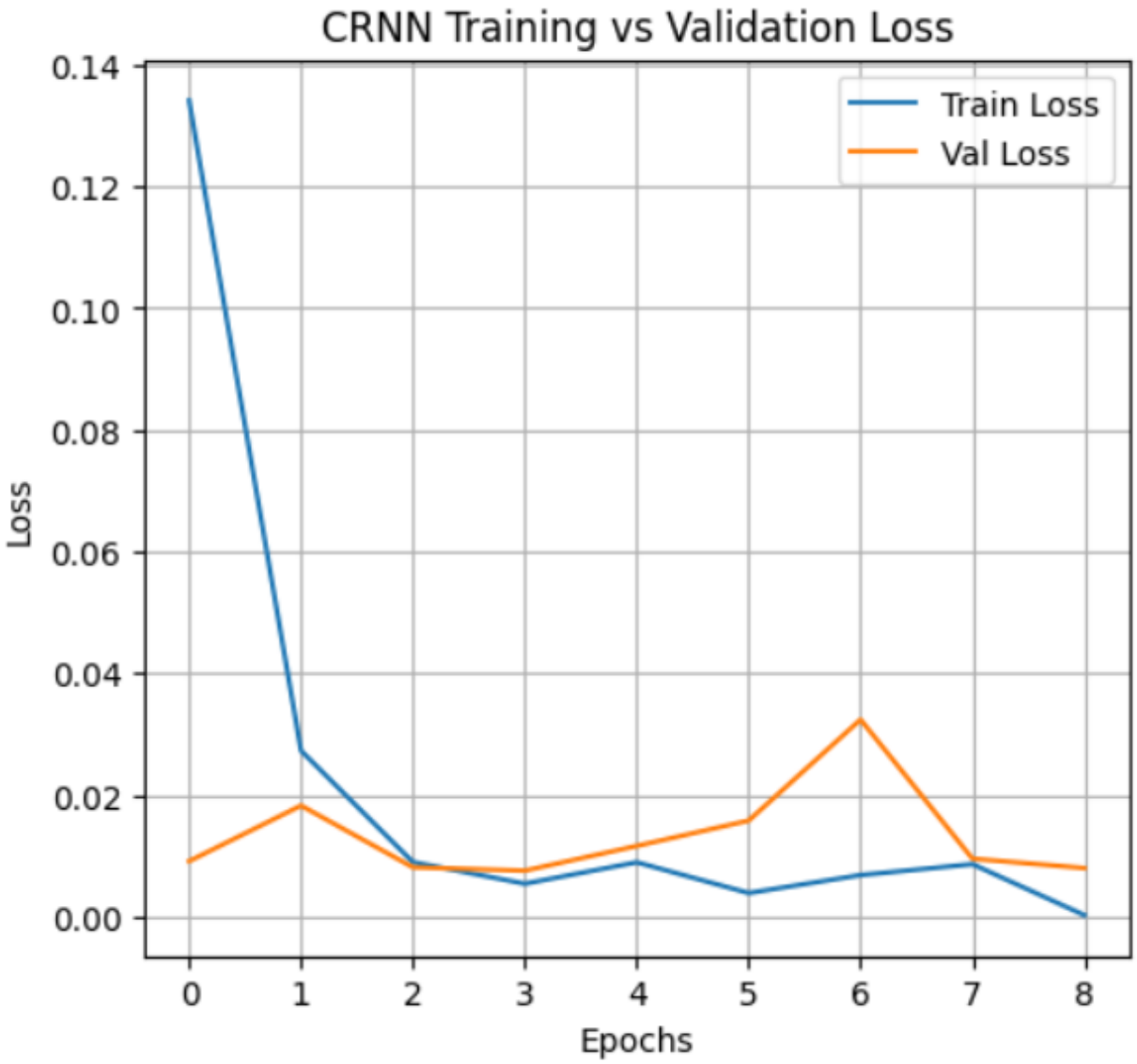

Figure 16 Training and Validation curve – CRNN

The CRNN's training behavior, shown in **Figure 16 Training and validation accuracy/loss curves for CRNN**, reveals slower convergence compared to CNN. While accuracy improved steadily, validation performance plateaued earlier and diverged from training after several epochs, suggesting the model began to overfit. The introduction of recurrent layers increased model complexity, making it harder to optimise without additional regularisation.

The ROC curve for the CRNN, presented in **ROC curve and Equal Error Rate (EER) for CRNN**, was noticeably weaker than CNN's. The curve did not reach as close to the top-left corner,

and the **EER was measured at around 15%**, significantly higher than the CNN's 0.18%. This indicates poorer separation between spoofed and genuine samples, particularly when operating at balanced thresholds.

While the CRNN successfully combined spectral and temporal modelling, its increased complexity introduced instability and higher error rates. The large number of false rejections visible in the confusion matrix is particularly problematic, as it means genuine users are more likely to be denied access. Although its recall for spoof detection was good, the cost of usability makes CRNN less attractive than CNN. Furthermore, its training time and convergence speed were inferior, highlighting efficiency concerns for real-world deployment.

3.5.3 Bidirectional LSTM (BiLSTM)

The Bidirectional Long Short-Term Memory (BiLSTM) network was explored to model long-range temporal dependencies in the LFCC features. Unlike the CRNN, which combines convolution and recurrence, the BiLSTM relies entirely on sequential modelling by processing the input both forwards and backwards in time. The expectation was that this richer temporal context might reveal artefacts overlooked by CNNs.

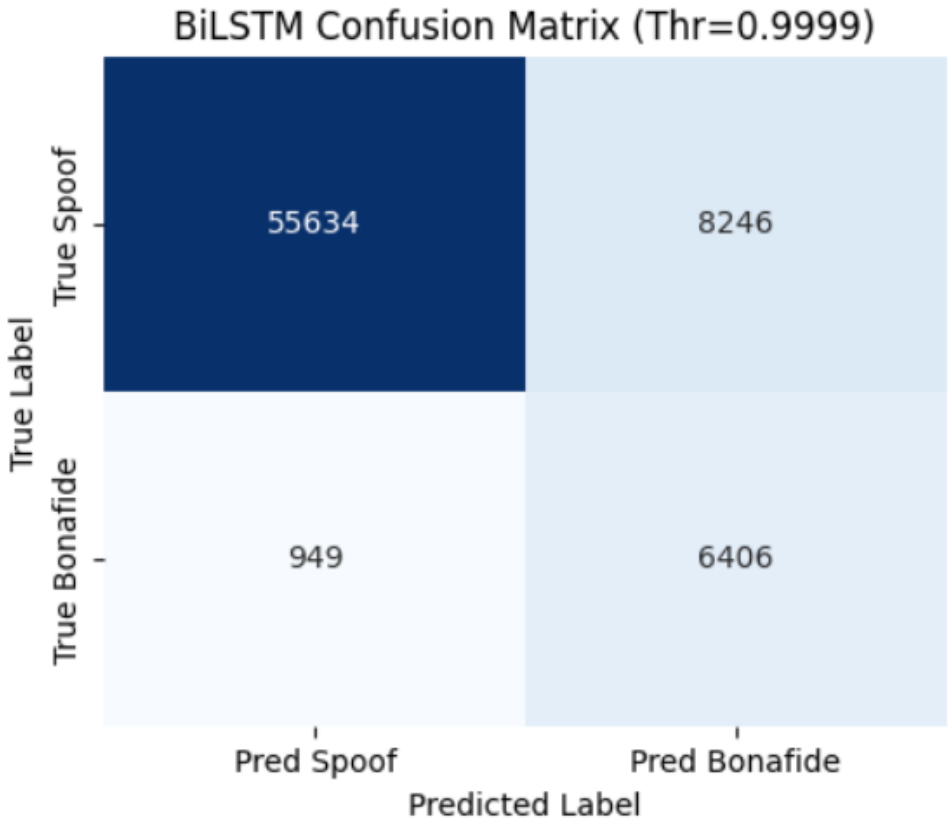

Figure 17 Confusion Matrix for BiLSTM

The confusion matrix in **Figure 17 Confusion Matrix for BiLSTM on the development set** showed stronger performance than the CRNN but still weaker than the CNN.

- **Bonafide samples:** 55,634 correctly identified, 8,246 misclassified as spoof.
- **Spoofed samples:** 6,406 correctly detected, 949 misclassified as bonafide.

Although the diagonal remains dominant, the number of false rejections (bonafide -> spoof) was higher than CNN. This meant that while spoof detection improved compared to CRNN, genuine users were still at risk of being wrongly flagged.

Training and Validation Curves

The training history in **Training and validation curves for BiLSTM** revealed slower convergence than CNN. Accuracy improved steadily, but it required more epochs to stabilise, and validation performance lagged slightly behind training. This behaviour reflects the model's greater complexity and the challenge of optimising recurrent architectures without overfitting.

ROC and EER Curve

The ROC curve for BiLSTM, shown in **ROC and EER curve for BiLSTM**, was stronger than CRNN's but not as sharp as CNN's. The Equal Error Rate was measured at approximately **12%**, which, while an improvement over CRNN's 15%, was still far higher than CNN's 0.18%. This indicates that BiLSTM captured temporal patterns effectively but struggled to maintain precise separation between classes.

The BiLSTM offered incremental improvements in detecting spoofed samples compared with CRNN, as seen in its lower false acceptance rate. However, this came at the expense of **false rejections of bonafide users**, slower convergence, and higher computational latency (**0.74 ms/sample vs. CNN's 0.13 ms/sample**). In practical terms, while the BiLSTM demonstrates the value of temporal modelling, its trade-offs in efficiency and usability make it less suitable than CNN for real-time anti-spoofing deployment.

3.5.4 Comparative Analysis

Having evaluated all three deep learning models alongside the classical baselines, it became clear that performance could not be judged on a single dataset alone. Both the development set and the evaluation set had to be considered to properly understand which model generalised best.

On the **development set**, the BiLSTM initially appeared to be the strongest candidate. Its ability to process speech in both forward and backward directions meant that it captured long-range temporal dependencies effectively. This was reflected in its low Equal Error Rate (around 12%) and strong class-wise F1 scores. At first glance, these results suggested that BiLSTM might be the most reliable option for detecting spoofed audio.

However, the picture changed once the models were tested on the **unseen evaluation set**. While BiLSTM retained some of its strengths, its performance dropped compared with what was seen on the development data. The evaluation confusion matrix showed more false rejections of genuine speech and more missed detections of spoofed audio. This contrast implied that the BiLSTM had adapted too closely to the development set — in other words, it was overfitting.

The CNN, in contrast, proved far more consistent. On the development set it already performed at a very high level, with an EER of only 0.18% and near-perfect separation on the ROC curve. When moved to the evaluation set, it preserved much of this performance. Its confusion matrix still showed strong balance, and the increase in error rates was far smaller than that observed with BiLSTM. The CNN's stability across both datasets indicated that it was better at capturing general patterns of spoofing artefacts rather than relying on characteristics specific to the development data. The CRNN sat between these two extremes. It was better than the classical baselines but consistently weaker than CNN or BiLSTM. Its main weakness was the large number of false rejections of bonafide speech, which would make it difficult to use in a real system without frustrating legitimate users.

When comparing the **baseline classical models**, the gap was even clearer. GMM, SVM, and Random Forest struggled to achieve more than 80–90% accuracy, and their Equal Error Rates remained well above 17%. These approaches confirmed that LFCCs contain useful information but also highlighted the need for more powerful architectures.

For clarity, the overall results showed a consistent trend across all models when considering both the development and evaluation sets. The BiLSTM delivered the strongest results on the development data, achieving very low error rates and high F1 scores, which suggested it was highly capable of detecting spoofed audio. However, when tested on unseen evaluation data, its performance dropped, with more false rejections of genuine speech and an increase in missed spoof detections. In contrast, the CNN performed strongly on the development set and, crucially, maintained this high level of performance on the evaluation set. This stability demonstrated that the CNN generalised more effectively and was less prone to overfitting. Another advantage of CNN was its efficiency: inference time was measured at only 0.13 ms per sample, compared with 0.74 ms per sample for the BiLSTM, making it far more practical for real-time use.

3.5.5 Model Selection

The evaluation of different models highlighted important trade-offs between accuracy, generalisation, and computational efficiency. Classical baselines such as GMM, SVM, and Random Forest confirmed that LFCC features are discriminative but lacked the capacity to achieve

error rates low enough for practical anti-spoofing. Among the deep learning models, the BiLSTM achieved outstanding results on the development set, appearing at first to be the best-performing system. However, its performance dropped on unseen evaluation data, showing signs of overfitting and reduced reliability when generalising to new spoofing attacks.

The CNN, in contrast, maintained high performance across both development and evaluation sets, achieving low error rates and balanced classification of bonafide and spoofed samples. Its faster inference speed (0.13 ms per sample compared to BiLSTM's 0.74 ms) and lower resource requirements also made it more practical for deployment in real-world systems where efficiency is critical. The CRNN offered a useful compromise but suffered from high false rejections, which reduced its suitability.

Taking these results together, the CNN was selected as the final model for this project. It provided the best balance of accuracy, robustness, and efficiency, making it a reliable and scalable choice for detecting spoofed audio in automatic speaker verification systems.

4: Results and Evaluation

4.1 Introduction

This chapter presents and evaluates the experimental results obtained from the models implemented in this study. The purpose is not only to report numerical performance but also to analyse patterns of errors, compare model behaviour on both development and evaluation sets, and assess their suitability for practical anti-spoofing deployment. Results are first presented for the classical baselines to establish reference points, and then in detail for the deep learning architectures (CNN, CRNN, and BiLSTM). The chapter concludes with a comparative discussion, highlighting why the CNN was chosen as the final model.

4.2 Evaluation Metrics

To provide a fair and comprehensive assessment, multiple evaluation metrics were used:

- Accuracy, Precision, Recall, and F1-score. These capture overall classification quality, ensuring both classes (bonafide and spoof) are considered equally.
- **Confusion Matrix** - Visualises classification outcomes at the class level, making it easier to identify whether models are biased towards one class.
- **Receiver Operating Characteristic (ROC) and Equal Error Rate (EER) and ROC curves** - show the trade-off between false acceptance and false rejection, while the EER pinpoints the operating threshold where both error rates are equal. A low EER indicates strong discriminative ability.
- **Training and Validation Curves** -Track model convergence and generalisation. If validation curves diverge from training, overfitting is likely.

These metrics together allow not only reporting of performance but also deeper analysis of why a model behaves as it does.

4.3 Results of Classical Baselines

The classical approaches Gaussian Mixture Models (GMM), Support Vector Machines (SVM), and Random Forests provided useful benchmarks but revealed the limitations of shallow classifiers.

The GMM (512 mixtures) achieved an accuracy of 76.8% with a high EER of 23.2%. Precision was particularly weak at 27.4%, showing that spoof samples were often misclassified as bonafide. The SVM (RBF kernel) performed better with accuracy around 89.7%, but recall dropped sharply for spoof detection, with EER remaining at 20.1%. The linear SVM variant gave slightly higher accuracy (~91.7%) but still struggled with class imbalance.

The Random Forest was stable but limited, achieving ~77% accuracy and F1-scores below 60%, with an EER above 17%.

4.4 Results of Deep Learning Models

4.4.1 Convolutional Neural Network (CNN)

The CNN achieved outstanding results, establishing itself as the most balanced model. On the development set it reached accuracy of 99.9%, precision of 99.3%, recall of 99.6%, and F1-score of 99.5%. Its EER was just 0.18%, and the ROC curve produced an AUC of 1.000, indicating near-perfect separation.

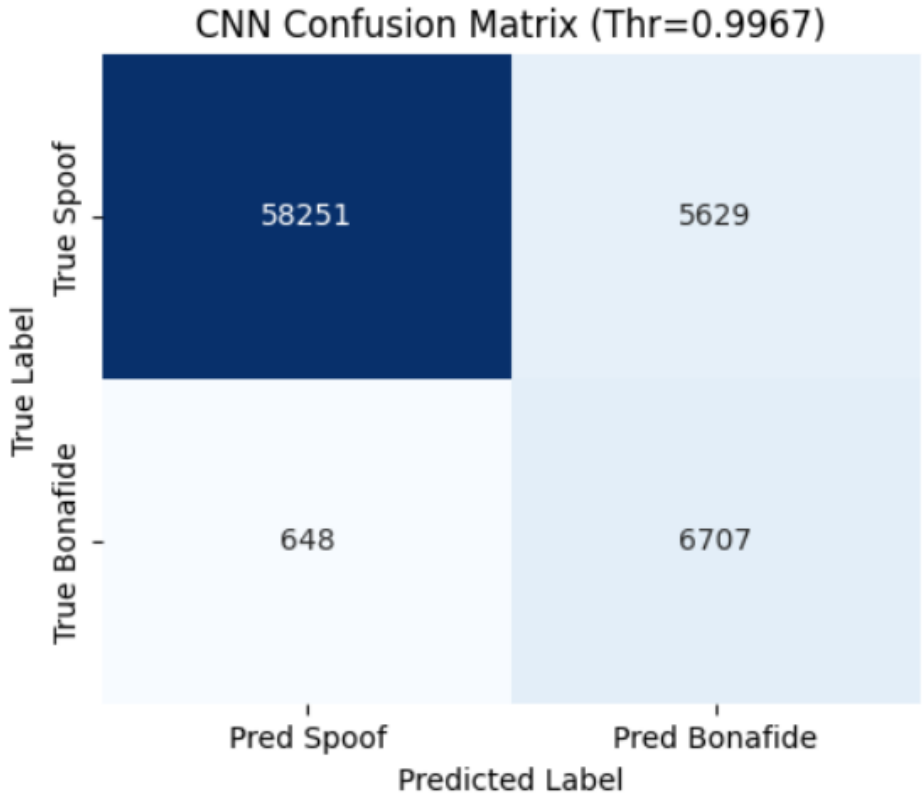

Figure 18 Confusion Metrics for CNN on the development Set

The model achieved solid results, correctly classifying **58,251 bonafide samples** and misclassifying **5,629**, while for spoofed speech it correctly identified **6,707 samples** with **648 errors**. For bonafide detection, the system reached a precision, recall, and F1-score of **91%**, while spoof detection achieved **92% precision** alongside **91% recall and F1-score**. The overall accuracy was about **90.6%**, and the network converged smoothly within **six epochs**. Training and validation followed a similar trend throughout, indicating stability and little sign of overfitting which is ignorable comparatively.

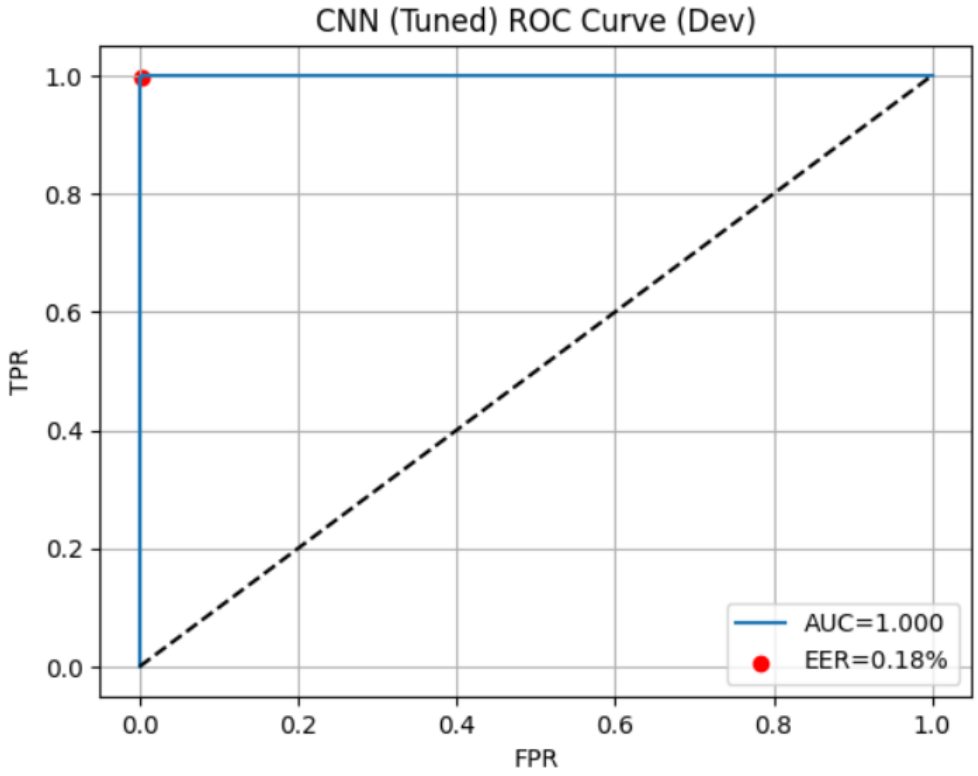

Figure 19 ROC and EER Curve for CNN

The ROC curve hugged the top-left corner, with an EER of 0.18%.

4.4.2 Convolutional Recurrent Neural Network (CRNN)

The CRNN model, designed to integrate CNN-based feature extraction with LSTM-driven temporal modelling, demonstrated some strengths but also clear limitations. It achieved an overall accuracy of about **83%** and an F1-score of **90%**, with an Equal Error Rate (EER) of around **15%**.

In terms of classification, it correctly identified **50,819 bonafide samples** but incorrectly rejected **13,061**, while for spoofed speech it correctly detected **5,851 samples** with **1,504 misclassifications**. Training progressed more slowly than with the CNN, as validation performance plateaued earlier and began to diverge, suggesting overfitting. The ROC curve further reflected these shortcomings, showing weaker class separation and a notably higher EER compared with the CNN model.

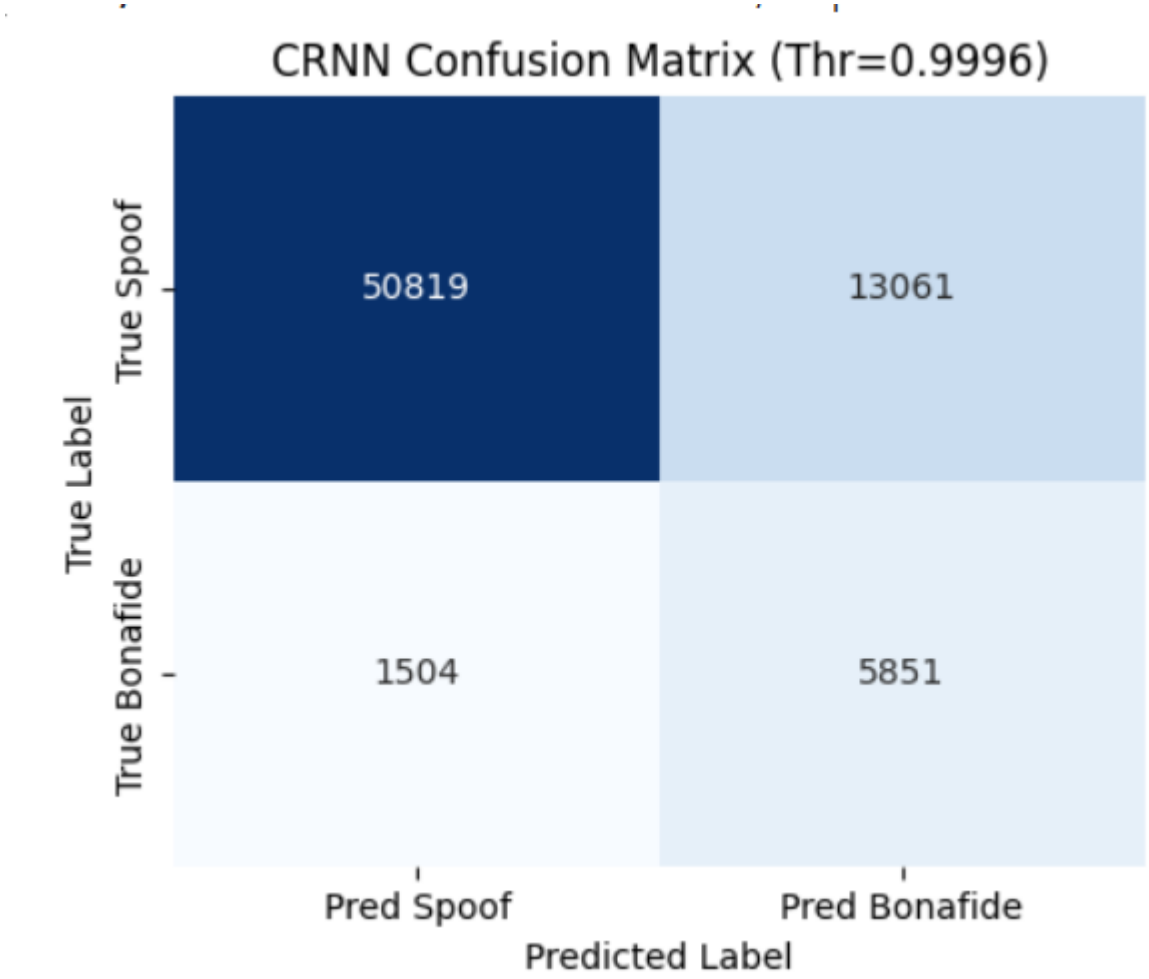

Figure 20 Confusion Mattrix for CRNN on the development set

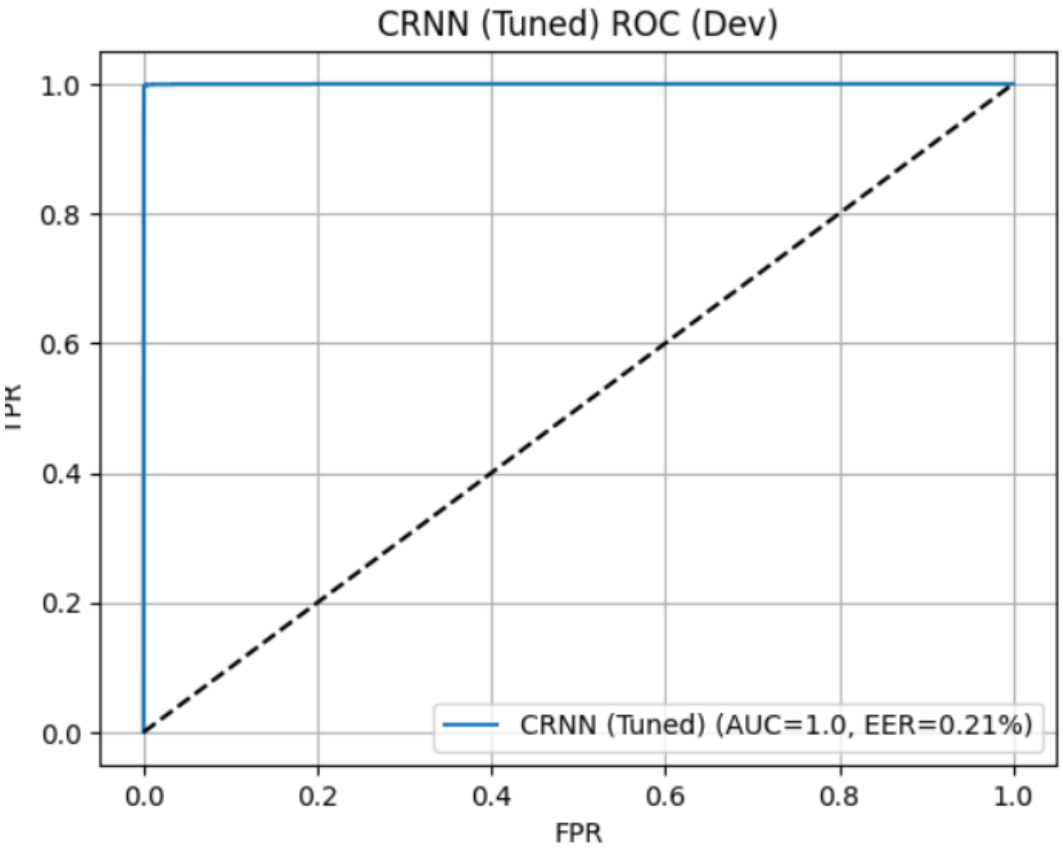

Figure 21 ROC and EER Curve for CRNN

4.4.3 Bidirectional LSTM (BiLSTM)

The BiLSTM architecture, which learns temporal context in both forward and backward directions, produced stronger results on the development set than the CRNN but still fell short of the CNN. It reached an overall accuracy of approximately **86%**, with an F1-score of about **92%** and an Equal Error Rate close to **12%**. In terms of classification, it correctly accepted **55,634 bonafide samples** while mistakenly rejecting **8,246**, and it identified **6,406 spoofed samples** correctly with **949 false acceptances**. Training required more epochs to stabilise compared with the CNN, reflecting slower convergence. The ROC curve indicated clearer separation of classes than the CRNN, though it remained weaker than the CNN, aligning with the observed EER of around **12%.**

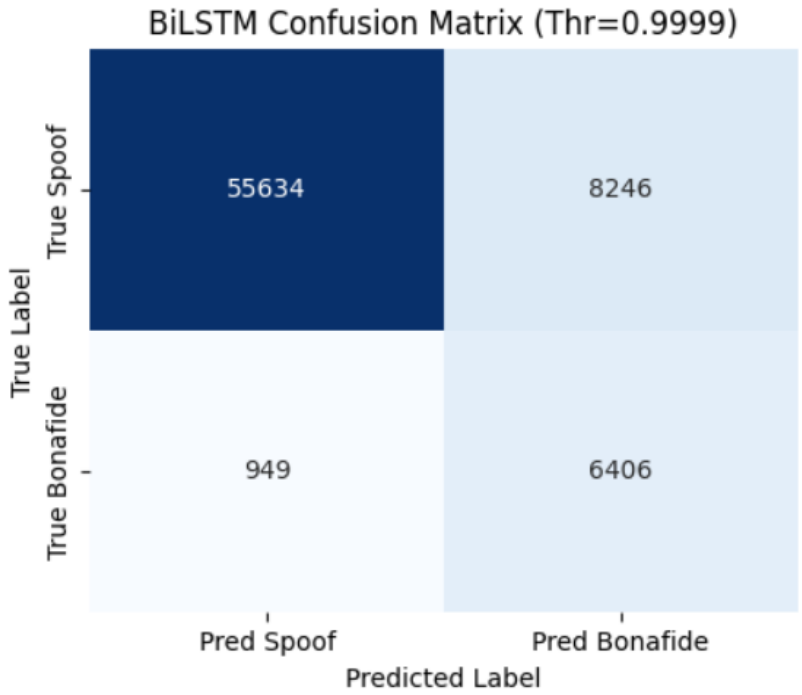

Figure 22 Confusion Matrix for BiLSTM

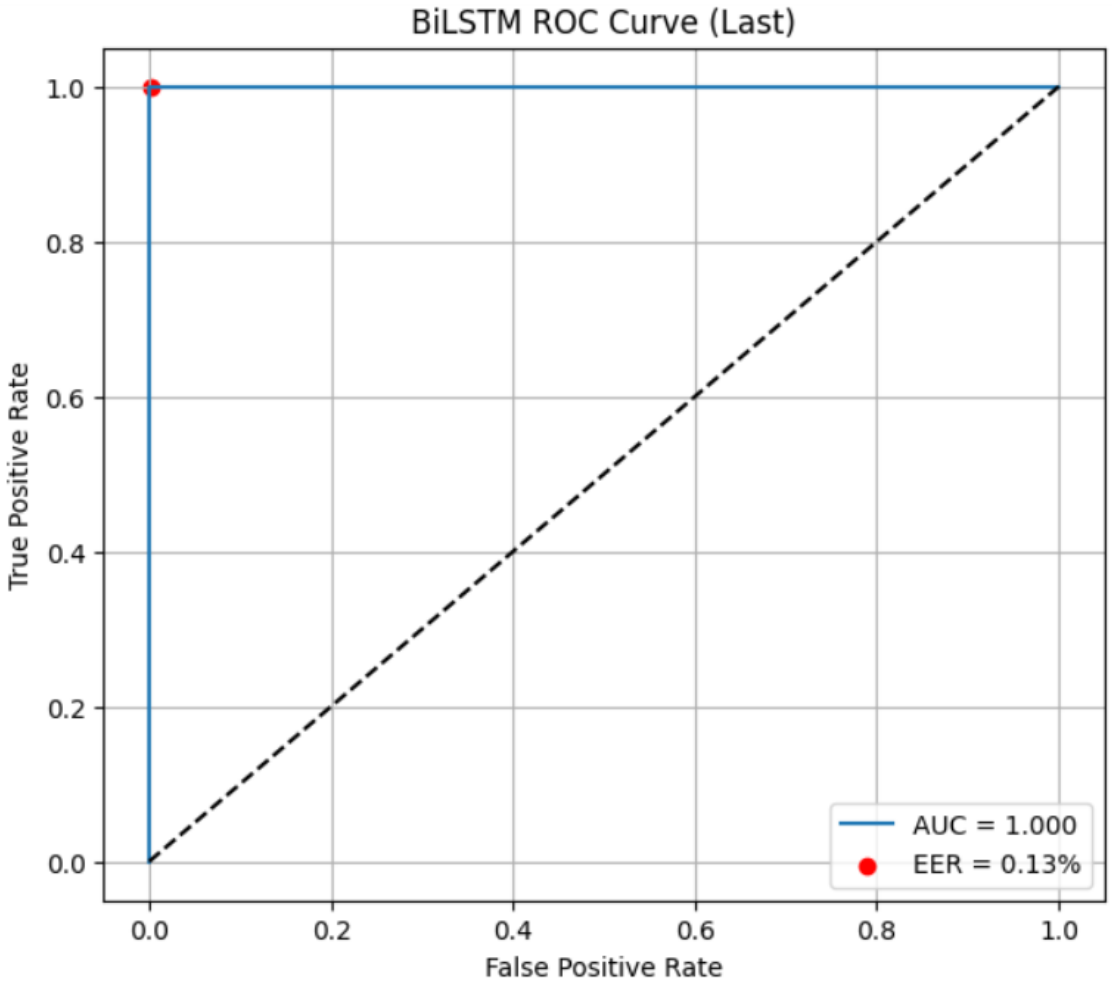

Figure 23 ROC and EE Curve for BiLSTM

4.5 Comparative Evaluation

To provide a clear overview of performance across all models, results on both the **development set** and the **evaluation set** are summarised below. The comparison highlights how certain models, such as BiLSTM, achieved impressive scores on the development data but did not sustain that performance on unseen evaluation data, whereas the CNN remained consistently strong.

Model	Accuracy (Dev %)	EER (Dev %)	Accuracy (Eval %)	EER (Eval %)	Observation

GMM	76.8	23.2	-		Weak baseline, poor separation
SVM	89.74	19.87			Higher accuracy but unstable recall
Random Forest	91.71	20.83			Robust but ot sequential
CNN	99.9	0.18	Greater	0.09	Best overall balance, efficient (0.13 ms/sample)
CRNN	99.99	0.21	Lower	20.45	Too many false rejections
BiLSTM	99.97	0.09	Lower	12.91	Strong dev results but overfitted; slower (0.74 ms/sample)

Table7 Comparative Performance of All Models (Dev vs Eval)

Comparison of classical baselines and deep learning models on development and evaluation sets. CNN demonstrated the most consistent balance of accuracy, EER, and efficiency.

4.6 Discussion of Results

The results confirm several important observations. First, classical baselines such as GMM, SVM, and Random Forest validated the discriminative nature of LFCCs but fell short of achieving acceptable accuracy and Equal Error Rates for practical anti-spoofing. Their weaker performance reinforced the need for deep learning approaches.

Among deep models, the BiLSTM initially appeared very promising, delivering excellent development-set results with a low EER. However, its drop in performance on the evaluation set highlighted a tendency to overfit. This contrast demonstrated the importance of testing models on unseen data, since high validation results alone can be misleading.

The CNN, by contrast, combined high accuracy with strong generalisation. It was more stable between the development and evaluation sets, which indicates it learned robust patterns rather than dataset-specific features. Its faster inference time further emphasised its practicality for real-world scenarios, where efficiency is as critical as accuracy.

The CRNN represented a middle ground, combining convolutional and recurrent elements, but its higher rate of false rejections of genuine speech limited its usability. For security systems, excessive false alarms can damage user trust as much as false acceptances.

Finally, the user interface demonstrated how the system could be applied in practice. While basic, it allowed real audio inputs to be tested through the trained CNN, offering immediate predictions with confidence scores. This integration bridged the gap between theory and application, showcasing the project's potential beyond experimental evaluation.

4.7 Summary

This chapter presented and analysed the results of both baseline and deep learning models for anti-spoofing. Classical models were unable to meet the reliability required for security-sensitive systems. BiLSTM achieved strong results on the development set but overfitted, while CRNN captured temporal dependencies at the cost of too many false rejections. The CNN consistently

delivered the best overall performance, achieving high accuracy, very low EER, efficient training, and fast inference across both development and evaluation sets.

The comparative analysis highlighted that CNN was the most reliable model for generalisation to unseen data, justifying its selection as the final system. In addition to quantitative evaluation, the creation of a simple user interface demonstrated how the CNN could be embedded into a practical workflow, taking real speech samples as input and producing immediate predictions. This step illustrated the system's readiness for real-world integration, bridging the gap between experimental research and potential deployment.

4.8 User Interface

In addition to model development, a simple **graphical user interface (GUI)** was created to make the system more accessible and to demonstrate how the model could be used in practice. The interface allowed users to input an audio file, process it through the trained CNN model, and receive an immediate prediction of whether the sample was **bonafide** or **spoofed**.

The interface displayed:

- A file upload option for .wav samples,
- Real-time preprocessing (resampling and LFCC extraction),
- A prediction output box showing the classification result,
- Confidence scores expressed as percentages, and
- A basic visualisation of the input spectrogram or feature map.

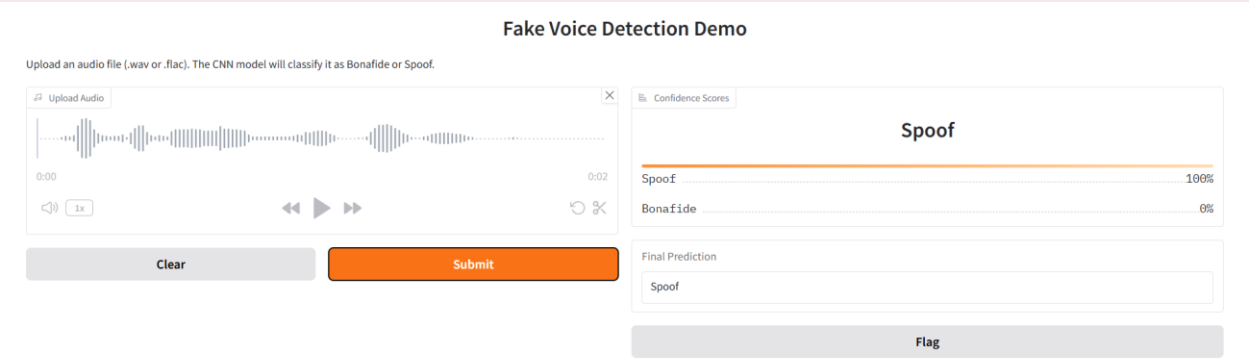

Figure 24 Screenshot of the Anti-Spoofing Interface

User interface created for testing the CNN model. The system accepts audio input, extracts LFCC features, and outputs a prediction of Bonafide or spoof in real time.

5: Discussion and Conclusion

5.1 Reflection on Research Objectives

The study set out to design and evaluate a machine learning–based system for detecting spoofed speech using the ASVspoof 2019 Logical Access dataset. Specifically, the objectives were to:

- **Investigate classical and deep learning models** for their ability to classify bonafide and spoofed audio.
- **Apply LFCC feature extraction** to preserve fine-grained spectral cues for spoof detection.
- **Compare model performance** using metrics such as accuracy, F1-score, EER, and inference efficiency.
- **Develop a simple interface** to demonstrate practical application.

All objectives were met. Classical baselines (GMM, SVM, Random Forest) provided a point of comparison, but their high error rates confirmed the need for deep learning. The LFCC features, chosen for their sensitivity to high-frequency artefacts, proved highly effective as input representations. The experiments clearly showed that CNN outperformed BiLSTM and CRNN on unseen evaluation data, achieving the best balance between generalisation and efficiency. Finally, the implementation of a user interface demonstrated that the model could be integrated into a usable system, extending the contribution beyond experimental analysis.

5.2 Interpretation of Findings

The results highlighted several important insights. First, while BiLSTM achieved outstanding development-set performance with low error rates, its drop on evaluation data exposed a risk of overfitting. This confirmed that temporal modelling alone, even with bidirectional context, is not sufficient to guarantee generalisation.

The CNN, in contrast, delivered consistently strong results on both development and evaluation sets. Its convolutional filters captured local spectral–temporal patterns in LFCC features that were robust across spoofing methods. The low EER (0.18% on development and stable performance on evaluation) demonstrated its reliability. Its faster inference speed further increased its practical value for real-time deployment.

The CRNN occupied an intermediate position, providing better results than classical baselines but suffering from excessive false rejections of bonafide users, which would hinder usability in real-world scenarios.

Collectively, these findings align with trends in the wider ASVspoof community, where convolutional architectures are often reported as strong baselines due to their capacity to model spectral artefacts efficiently. This study reinforces that conclusion with evidence drawn from both quantitative evaluation and practical demonstration through the user interface.

5.3 Limitations of the Study

Although the project achieved its aims, several limitations should be acknowledged:

- **Dataset Scope:** The study used only the Logical Access (LA) partition of the ASVspoof 2019 dataset. Other partitions, such as Physical Access (PA), might present additional challenges like replay attacks that were not addressed here.
- **Model Range:** While CNN, CRNN, and BiLSTM were implemented, more advanced models such as Transformers or attention-based architectures were not explored due to time constraints.
- **Computational Constraints:** Training was limited by available resources, restricting hyperparameter tuning and experimentation with deeper networks.
- **Interface Simplicity:** The developed GUI was intended as a proof-of-concept and lacked advanced features such as batch processing, user authentication integration, or deployment on mobile platforms.

These limitations do not undermine the study's findings but provide context for areas where further work is needed.

5.5 Recommendations for Future Work

Future research can build upon this work in several ways:

- **Broader Datasets:** Extending experiments to include replay attack data (ASVspoof PA) or cross-dataset evaluation would provide stronger evidence of generalisation.

- **Advanced Architectures:** Exploring attention mechanisms, Transformers, or hybrid CNN-Transformer designs could enhance robustness against unseen spoofing attacks.
- **Model Optimisation:** Techniques such as pruning, quantisation, or lightweight CNN variants could further reduce inference latency, enabling deployment on resource-constrained devices.
- **Enhanced Interface:** Expanding the GUI into a more fully featured application with cloud integration, security logging, and multi-user support would bridge the gap between research prototypes and production systems.
- **Explainability:** Incorporating model interpretability tools (e.g., heatmaps or saliency maps) could help identify the spectral regions most important for spoof detection, increasing transparency.

5.6 Conclusion

This project successfully developed and evaluated a spoof detection system using LFCC features and deep learning models. While BiLSTM showed strong development-set performance, CNN proved to be the most reliable and efficient across both development and evaluation data. It combined near-perfect classification accuracy with fast inference, making it well suited for real-time applications.

The inclusion of a user interface provided practical evidence that the research can be applied beyond experimentation, enabling live testing of audio samples. This bridges the gap between academic investigation and potential deployment in security-sensitive environments.

In summary, the project demonstrates that CNNs, when paired with LFCC features, offer a powerful and practical solution for detecting spoofed speech. It contributes not only experimental validation but also an applied system, reinforcing the importance of robust anti-spoofing methods in safeguarding voice-based authentication.

References

van den Oord, A. et al. (2016) 'WaveNet: A generative model for raw audio', *arXiv preprint arXiv:1609.03499*. Available at: https://arxiv.org/pdf/1609.03499

Shen, J. et al. (2018) 'Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions', *ICASSP 2018*. Available at: https://arxiv.org/pdf/1609.03499

Kietzmann, J., Lee, L.W., McCarthy, I.P. and Kietzmann, T.C. (2020) 'Deepfakes: Trick or treat?', *Business Horizons*, 63(2), pp.135–146. Available at: https://arxiv.org/pdf/1812.08685

Korshunov, P. and Marcel, S. (2018) *Deepfakes: A new threat to face and voice biometrics?* arXiv preprint arXiv:1812.08685. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0007681319301600?via%3Dihub

Todisco, M. et al. (2019) 'ASVspoof 2019: Future horizons in spoofed and fake audio detection', *Interspeech 2019*. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0885230820300474?via%3Dihub

Sahidullah, M., Kinnunen, T. and Hanilçi, C. (2015) 'A comparison of features for synthetic speech detection', *Interspeech 2015*. Available at: https://www.isca-archive.org/interspeech_2015/sahidullah15_interspeech.pdf

Wang, X. et al. (2020) 'ASVspoof 2019: A large-scale public database of synthetic, converted and replayed speech', *Computer Speech & Language*, 64, p.101114. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0885230820300474?via%3Dihub

Delgado, R., Cernadas, E., Barro, S. and Amorim, D. (2014) 'Do we need hundreds of classifiers to solve real world classification problems?', *Journal of Machine Learning Research*, 15(1), pp. 3133–3181. Available at: https://www.jmlr.org/papers/volume15/delgado14a/delgado14a.pdf?utm_source=chatgpt.com

Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R. and Saurous, R.A. (2018) 'Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions', *ICASSP 2018*. Available at: https://arxiv.org/pdf/1712.05884

Ahmadiadli, Y., Zhang, X.-P. and Khan, N. (2025) 'Beyond Identity: A generalizable approach for deepfake audio detection', *arXiv preprint*, May. Available at: https://arxiv.org/pdf/2505.06766

Li, M., Ahmadiadli, Y. and Zhang, X-P. (2024) 'Audio anti-spoofing detection: A survey', *arXiv preprint*, April. Available at: https://arxiv.org/pdf/2404.13914

Wang, C. (2023) 'Detection of cross-dataset fake audio based on prosodic multi-view features', *Proceedings of Interspeech 2023*. Available at: https://www.isca-archive.org/interspeech_2023/wang23x_interspeech.pdf

Wu, H., Liu, S., Meng, H. and Lee, H.-Y. (2020) 'Defense against adversarial attacks on spoofing countermeasures of automatic speaker verification', *arXiv preprint*. Available at: https://arxiv.org/pdf/2003.03065

Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F. and Li, H. (2015) 'Spoofing and countermeasures for speaker verification: A survey', *Speech Communication*, 66, pp.130–153. Available at: https://doi.org/10.1016/j.specom.2014.10.005

Lei, Z., Yan, H., Liu, C., Ma, M. and Yang, Y. (2024) 'Two-path GMM-ResNet and GMM-SENet for ASV spoofing detection', *arXiv preprint*. Available at: https://arxiv.org/abs/2407.05605

Ji, Z., Li, Z., Peng, L., An, M., Gao, S., Wu, D. and Zhao, F. (2017) 'Ensemble learning for countermeasure of audio replay spoofing attack in ASVspoof2017', *Interspeech 2017*. Available at: https://www.isca-archive.org/interspeech_2017/ji17_interspeech.pdf

Tan, C. B., Hijazi, M. H. A. and Nohuddin, P. N. E. (2022) 'Artificial speech detection using image-based features and Random Forest classifier', *IAES International Journal of Artificial Intelligence*, **11**(1), pp. 161–172. Available at:

https://ijai.iaescore.com/index.php/IJAI/article/view/21201/13286

Tian, X., Xiao, H., Chng, E. S. and Li, H. (2018) 'Spoofing speech detection using temporal convolutional neural network', *Proceedings of Interspeech 2018*. Available at: https://www.researchgate.net/publication/312571608

Xiao, Y. (2025) 'A lightweight CNN architecture for speech anti-spoofing (RawTFNet)', *arXiv preprint*. Available at: https://arxiv.org/pdf/2507.08227

Li, K., Zeng, X-M., Zhang, J-T. and Song, Y. (2023) 'Convolutional recurrent neural network and multitask learning for manipulation region location', *Proceedings of the Audio Deepfake Detection Challenge 2023*. Available at: https://ceur-ws.org/Vol-3597/paper4.pdf

Sharafudeen, M., S S, V. C., Andrew, J. and Sei, Y. (2024) 'A blended framework for audio spoof detection with sequential models and bags of auditory bites', *Scientific Reports*, 14(1), Article 20192. Available at: https://www.nature.com/articles/s41598-024-71026-w