Abstract

Automatic speech verification systems are currently seeing widespread demands in security-sensitive applications such as biometric authentication, access control, and remote identity verification. Whereas machine-learning-based methods have allowed these kinds of systems to become more effective at identifying forms of spoofing, studies have revealed that these systems remain highly susceptible to the forms of adversarial attack whereby minor and deliberately created changes in an audio sample can lead the system to make incorrect classifications of what it hears. Majority of studies that discuss the resilience of systems to adversarial attacks have centered on convolutional neural networks (CNNs); more recent studies in the last three years have indicated that countermeasures to CNN-based spoofing are not resilient enough by themselves to provide robustness especially when adversarial or mismatched conditions occur. Also, little has been done on lightweight countermeasures against spoofing which consume less computational resources and memory and will be used in resource-constrained settings. The thesis is that a lightweight log-mel-based feed-forward neural network is capable of defense against adversarial attacks to the ability to detect synthetic audio recordings. The network was compared to both white-box Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) adversarial attack algorithms and tested performance was compared with a classical CNN-based spoofing countermeasure that was designed to accomplish the same. The vulnerability of the network to these kinds of attacks was tested by using both structured ASVspoof2019 LA data set and an unconstrained real-world WILD data set. Findings reveal that, though the lightweight feed-forward network was rather robust in terms of baseline performance, it was even worse off than the more conventional CNN-based spoofing countermeasures when the feed-forward network was exposed to adversarially-generated noise. Thus, despite the classification accuracy of a countermeasure being very large, it is not ensured that the countermeasure is able to respond to synthetic audio recordings regardless of the circumstances, and in cases where developing deployable ASV countermeasures, it is unquestionably necessary to perform adversarial-aware testing.

1 Introduction

There has been an increase in the implementation of automatic speech verification (ASV) systems in security-sensitive systems like biometric authentication, access control, and remote identity verification. These advances in model based on machine-learning have led to their adoption by enhancing the fit between bona fide speech and spoofed or synthetic audio, which has increased dramatically. With the increasing penetration of ASV systems into actual security systems in the real world, the question of their reliability in adversarial and non-ideal conditions has become a major issue of concern.

Although it was showing good results in a typical assessment environment, recent research has established that ASV spoofing countermeasures are very susceptible to adversarial attacks. In such attacks, a system can be fooled into classifying synthetic speech as natural or the other way round due to imperceptible alterations that are introduced to an audio waveform. This security risk is critical, because the adversarial audio can be used to pass through the authentication procedures undetected by the system or human auditors. Most of the literature available on the topic of adversarial robustness in speech processing has been performed in the context of convolutional neural network (CNN) architectures because of the high performance observed when spectralgram-based representations are used.

Nevertheless, recent published literature has indicated that CNN-based spoofing mitigation measures are not adequate to guarantee robustness, especially in adversarial or misaligned testing situations. These models frequently are based on high-capacity architectures, which are sensitive to small input perturbations and are not well-behaved at generalization on uncontrolled datasets. Simultaneously, there has been little focus on the lightweight spoofing countermeasures, which can be implemented in the environment that is limited in resources, e.g. embedded devices or edge-based authentication systems. Such environments have hard constraints on computational complexity, memory consumption and latency, rendering most current adversarial defense strategies in impractical situations.

These gaps are filled in this thesis, and the study delves into the adversarial robustness of a lightweight feed-forward neural network to detect synthetic speech using log-mel spectral features. The suggested model will reduce the amount of computational and memory used and still compete on a level with other baselines. It is tested on the white-box adversarial threat models of the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) attacks. The model performance is compared to a conventional CNN-based spoofing countermeasure, which was created in the same classification task.

The Automatic Speaker Verification (ASV) Systems qualify to be a biometric authentication system of the end-users since it offers Non-Invasive Authentication to end-users; it is easy to use and may be introduced through an ASV system or a Telephone Banking Service, etc. Nevertheless, ASV systems can be attacked by a Spoofing attack where the person is trying to gain unauthorised access by impersonating an ASV system using a Replayed, Synthesised or Converted speech. The analysis of the ASV systems benchmark tests and surveys over the last couple of years revealed that Spoofing will remain an issue with ASV systems and that with the new detection techniques at the disposal, Spoofing can be a challenging issue to detect (Cai 2017; Kinnunen and Sahidullah 2017).

Attaining the solutions to address the menaces of the Spoofing in ASV Systems involve numerous Countermeasures; the majority of the suggested Countermeasures are founded on the model of the Machine Learning with Spectral Representations of speech, predominantly the Log-Mel Spectrograms. Though there are many different types of factors, the performance of most Machine Learning Methods on Benchmark Datasets has been demonstrated to be very high, the emerging issues with regard to how Robust these methods are to Adversarial Conditions has been brought up, the most recent studies have indicated that the Countermeasures might be Manipulated by introducing Adversarial Perturbations to the input to the Countermeasures to induce Misclassification of input (Liu et al 2019).

The threat model of ASV systems posed by Adversarial Machine Learning and audio-based biometric systems is very real. Specifically, gradient-based attacks, including Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) utilize sensitivity of the neural networks to input perturbations that cause a maximum classification error (Goodfellow et al. 2015; Madry et al. 2018). It has been demonstrated that these attacks have a dramatic impact on the performance of all ASV systems even in cases where the perturbations are not perceptually salient.

The effect of participants of such attacks is a false acceptance rate (FAR) which increases the false acceptance rate (FAR) of such types of attacks, which is of particular concern since the adversarial attack in ASV attacks can lead to an increase in false acceptance rate. According to Liu et al. (2019), several state-of-the-art countermeasures that are based on convolutional neural network (CNN) have experienced disastrous losses in performance under both FGSM

and PGD attacks. It is also established by other researchers that adversarial vulnerability seems to be a universal problem in the architecture of ASV and feature representation (Wu et al. 2020). Automatic speaker verification (ASV) systems are becoming more popular in security sensitive applications like biometric authentication, access control and remote identity verification. Since these systems are implemented at scale, the vulnerability to malicious manipulation is an essential issue. Even though recent machine-learning-based spoofing countermeasures have been demonstrated to show high classification accuracy in benign settings, there is increasing evidence that the reliability of these countermeasures substantially reduces when they have to deal with adversarially manipulated speech. This casts grave doubts on the viability of existing ASV systems in practice in security environments.

The ASV adversarial attacks are particularly concerning due to their main effect that raises the false acceptance rate (FAR), that is, false knowledge of spoofed or synthetic audio as authentic. As demonstrated by Liu et al. (2019), a number of state-of-the-art convolutional neural network (CNN)-based spoofing countermeasures are catastrophic to Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) attacks. These results have been supported later by other researchers showing that adversarial vulnerability is a general problem in ASV models, methods of training, and feature representations and not a limited feature of individual models (Wu et al., 2020; Wu et al., 2022).

The majority of the prevailing literature on adversarial robustness in ASV concerns deep CNN models that are trained on organized benchmark datasets like ASVspoof2019 LA. These datasets offer a controlled and standardized assessment structure, although they do not entirely represent the variety and uncertainty of the real-world audio conditions, such as environmental noise, channels variability, and artifacts of recordings (Kinnunen and Sahidullah, 2017). The strong conclusions made based on structured datasets, therefore, might not be applicable to non-constrained deployment cases. Moreover, deep CNN-based countermeasures can be computationally intensive and thus can be problematic to implement in settings where efficiency and low latency are required.

Feed-forward neural networks and spectral features are used to create lightweight spoofing countermeasures, which are a viable option in resource-constrained environments. Although they are relevant, there is comparatively little literature that studies their strength against adversarial attacks. It is also not yet clear that the resistance to adversarial perturbations to increasing model complexity is inherent or that adversarial vulnerability is a structural feature of learning-based spoofing defenses, irrespective of the model depth (Wu et al., 2020; Wu et al., 2022). This knowledge gap in the literature will inform a dedicated study of the adversarial nature of lightweight ASV countermeasures.

This thesis examines the strength of a colorblind log-mel-based feed-forward neural network to white-box adversarial attacks that operate on the feature-level. The suggested model is tested on the FGSM and PGD attacks and compared to a standard CNN-based spoofing countermeasure, which also performs the same classification task. The structured ASVspoof2019 LA is also experimented on an unconstrained real-world WILD dataset to evaluate the vulnerability by dataset. The work makes such contributions as a systematic analysis of the adversarial resilience of lightweight ASV countermeasures, direct comparison against CNN-based methods, and the evaluation of the difference between controlled and natural-world audio settings and adversarial vulnerability. Collectively, these results show how much accuracy-based assessment is restricted and the importance of adversarial-sensitive testing in the design of deployable ASV systems.

The key contributions of this work are:

- An adversarial robustness evaluation of a lightweight spoofing countermeasure
- A comparative analysis between lightweight and CNN-based models
- A dataset-level comparison highlighting differences between benchmark and real-world conditions
- Empirical evidence supporting prior findings on the fragility of spoofing countermeasures under adversarial attacks (Liu et al., 2019; Wu et al., 2020)

2 Literature Review

It has been established that Automatic Speaker Verification (ASV) systems are vulnerable to spoofing by more advanced applications in voice conversion and voice replay attacks developed in the recent past (Kinnunen and Sahidullah 2017). These systems were not generalisable to a variety of different forms of spoofing and the traditional countermeasures relied on hand crafted features and shallow classifiers. The ASVspoof challenges revealed the weakness of these countermeasures, and more advanced anti-spoofing technologies on deep learning were created.

Recently a new line of thinking with regard to intelligent systems has emerged through the use of deep learning as a whole. Such systems incorporate an expanding pool of convolutional neural network (CNN) systems and spectral information deep learning systems. As an example, Salih et al., 2025 have shown that it was effective to use CNN trained on the spectrogram and the mel frequencies to identify the presence of a spoofing attack in the context of automatic speaker verification system (ASV). Most ASVspoof systems are based on CNN architecture (LCNN-big, LCNN-small and SENet12).

However, the strength of these deep models is counterbalanced by three issues:

1. **Overfitting to dataset-specific artefacts**, limiting generalisation.
2. **High computational cost**, making them unsuitable for low-resource deployments.
3. **Documented fragility to adversarial perturbations**, which dramatically distort performance.

The findings of a study by Liu, Wu, Lee, and Meng (2019) established the fact that the inherent security concerns of many speech-processing models have been proven by the highly influential vulnerabilities of the adversarial ML that the researchers identified in their study. It was also revealed in the study that CNNs are effective, when tested by the use of FGSM and PGD methods, and led to the performance degradation of the spoofing countermeasures even at extremely low perturbation levels. The research demonstrates that most of the anti-spoofing systems perform well do not necessarily imply that they would be secure.

This list of vulnerabilities is confirmed and extended to the analysis of the two types of attacks white-box and transfer attacks (Wu et al., 2020). Their investigation has confirmed that adversarial perturbations worked well across the different forms of speech-processing systems and that they did not rise with the complexity of the model; therefore, the modelling of more complex neural networks was not linked to ensuring more significant levels of adversarial robustness (Kassis and Hengartner, 2021).

More sophisticated techniques have been developed to generate adversarial attack such as Malafide, which is a universal connotative noise attack developed by Panariello et al. (2023) that does not need access to gradient maps of utterances; thus, making it work on any utterance style. As Kassis and Hengartner (2021) show, adversarial perturbation generated on the basis of physical-world sounds can be generated in real-time, allowing one to use the practice sample of attack settings that are not confined to controlled/laboratory situations.

These papers reinforce the fact that the adversarial attacks are a real and very practical threat to the ASV systems.

Defences against adversarial attacks typically fall into two categories:

Wu et al. (2020) demonstrated that adversarial training can enhance robustness at the cost of being computationally expensive and potentially lowering clean-speech performance. Even though the above-mentioned mitigation techniques are not so robust as many other more sophisticated and heavier methods of gradient descent, the abovementioned mitigation techniques still provide certain protection against less powerful attacks by using self-supervised feature learning (Wu et al., 2022). Besides using high numbers of computers to train and test their results, self-supervised feature learning methods are yet to provide a route towards the study of lighter systems. According to Khan, Malik & Ryan (2022), lightweight models with log-mel or cepstral features that utilize Multi-Layer Perceptrons (MLPs) or other small models (in terms of depth) are generally adopted because of their high processing speed and little power demands. Most lightweight models have been evaluated against adversarial attack but there is a dearth of models tested against their robustness to adversarial attack. It is believed that due to the reduced complexity of such models (i.e. the reduced decision boundaries) and due to limited representations of their inputs, such models can be less resistant to adversarial attacks compared to CNNs but this remains yet to be verified; hence a significant field of interest that needs further research.

Modern adversarial attacks increasingly account for real-world constraints:

- **Universal filters** (Panariello et al., 2023) require no per-sample optimisation.
- **Time-domain perturbations** (Kassis & Hengartner, 2021) make physical deployment feasible.
- **Perturbation transferability** enables black-box attacks without model access.

These developments show that adversarial threats are no longer theoretical—they can be deployed practically against real ASV systems. This further amplifies the need to understand the vulnerabilities of lightweight models.

Across the reviewed literature, several critical gaps emerge:

1. **Lightweight log-mel–based models remain untested against adversarial attacks**, despite being common in deployed ASV systems.
2. **Most studies evaluate robustness only on structured datasets**, neglecting real-world acoustic variability.
3. **Comparative robustness between lightweight and deep models is unknown**.
4. **Existing defences focus primarily on CNNs**, offering no guidance for resource-constrained environments.

The present research offers a contribution to addressing these gaps in knowledge in the domain of adversarial robustness, namely; (a) the performance of the lightweight MLP model and powerful adversarial attacks, (b) the comparisons of the MLP and CNNs trained on the ASVspoof and WILD datasets, and (c) the evaluations of the CNN models that have been reproduced and benchmarked to compare with those of Liu et al. (2019).

This section is, then, aimed at synthesizing the key outputs that came about when this project was being executed, output like ready datasets to be experimented with, adopted models, adversarial attack processes, assessment artefacts, and analysis results. All the outputs substantially direct a part of the general trends of the objectives of the research expressed earlier.

In more recent developments, the dominance of convolutional neural network (CNN) architectures trained on spectral representations of speech, including log-mel spectrograms, has taken the place of the dominance of convolutional neural network (CNN) architectures on image-based data. The main components of many current state-of-the-art ASVspoof countermeasures rely on CNN-based systems such as LCNN and SENet variants and have demonstrated a high level of performance on benchmark datasets (Salih et al., 2025). Nevertheless, these achievements have been accompanied by a number of studies that determine the main disadvantages of deep models, such as their vulnerability to dataset-specific artefacts, which make them costly and unsuitable in low-resource or latency-sensitive deployment settings.

In addition to the increase in the detection accuracy of spoofing, adversarial machine learning has also shown the existence of basic security vulnerabilities in the countermeasures to ASV based on learning. The first systematic study which proved that CNN-based spoofing defenses are sensitive to gradient-based adversarial attacks including Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) was by Liu et al. (2019). Their findings revealed that the slight and insignificant modifications can lead to disastrous loss of performance and create a high rate of false acceptance. Later researchers found that adversarial vulnerability represents a common issue across ASV models and feature representations and that more complex models do not necessarily exhibit greater robustness (Wu et al., 2020; Kassis and Hengartner, 2021).

Recent literature has generalized adversarial threat models to white-box attacks. Panariello et al. (2023) came up with a universal adversarial attack named Malafide, which does not need access to model gradients and is also applicable to any utterance. Simultaneously, Kassis and Hengartner (2021) showed that adversarial perturbations could be produced in the time domain, and put into the real-world acoustic setting, which also contributes to the plausibility of physical-world attacks. Those developments show that adversarial threats to ASV systems are not just theoretical anymore but are actual threats to deployed systems.

There are a number of defence mechanisms that are suggested to address adversarial attacks on ASV countermeasures. Adversarial training is demonstrated to enhance robustness to certain types of attacks, but at a high computational cost and it affects clean-speech performance (Wu et al., 2020). Later methods using self-supervised learning have also exhibited better adversarial resilience, but usually large-scale models and significant amounts of computing power, which restricts their use to lightweight systems (Wu et al., 2022). Consequently, the current defence mechanisms are based mostly on deep architectures and give minimal guidance to resource-constrained deployments.

Lightweight spoofing controls and the log-mel or cepstral features coupled with feed-forward neural network or Multi-layer perceptron have also been particularly popular in real-world

applications of ASV because of their efficiency and power requirements (Khan, Malik and Ryan, 2022). Regardless of their prevalence, comparatively poor studies have been conducted to evaluate their resilience to adversarial attacks. The question of whether lower-complexity architectures produce higher adversarial vulnerability or that adversarial vulnerability is a common characteristic of learning-based spoofing detection systems regardless of their model depth is open.

There are a number of gaps that can be identified across the literature that is in place. First, the lightweight log-mel-based countermeasures have never been thoroughly tested against the strong adversarial attacks. Second, the vast majority of robustness studies are only based on structured benchmark datasets, ignoring the variation in recording conditions in the real world. Third, there are limited direct comparisons of lightweight and deep countermeasures in the same adversarial conditions. These gaps are critical to the interpretation of security implication of application of lightweight ASV spoofing countermeasures in actual settings.

Output Name	Type	Description	Primary Purpose	Reference
ASVspoof2019 LA Processed Dataset	Dataset	Structured benchmark dataset prepared using official ASVspoof protocol files, containing labelled bona fide and spoofed samples with extracted spectral features.	Baseline and adversarial evaluation under controlled benchmark conditions.	Kinnunen & Sahidullah (2017)
WILD Processed Dataset	Dataset	Real-world audio dataset processed into log-mel representations with consistent formatting and normalisation.	Evaluation under unconstrained, real-world conditions.	This work
Lightweight Log-Mel MLP Model	Model	Feed-forward neural network trained on flattened log-mel spectrogram features for spoofing detection.	Primary model under adversarial robustness evaluation.	This work
Reproduced CNN Countermeasure Models	Model	Reimplementation of CNN-based spoofing countermeasures following the architectures described by Liu et al. (2019).	Comparative adversarial robustness benchmarking.	Liu et al. (2019)
FGSM Attack Module	Algorithm / Code	Implementation of the Fast Gradient	Evaluation of first-order	Goodfellow et al. (2015)

		Sign Method for generating adversarial examples using first-order gradient information.	adversarial vulnerability.	
PGD Attack Module	Algorithm / Code	Iterative Projected Gradient Descent attack with configurable parameters to approximate worst-case adversarial behaviour.	Evaluation under strong adversarial attack conditions.	Madry et al. (2018)
Evaluation and Metrics Framework	Tool	Scripts to compute accuracy, FAR, FRR, EER, and attack success rate on clean and adversarial inputs.	Quantitative performance and security assessment.	Liu et al. (2019)
Visual Analysis Outputs	Figures	Confusion matrices, performance degradation plots, and comparative charts.	Interpretation and presentation of experimental results.	This work

Output Name	**Type**	**Description**	**Primary Purpose**

ASVspoof2019 LA Processed Dataset	Dataset	Structured dataset prepared using official protocol files, containing labelled bona fide and spoofed samples with extracted spectral features.	Baseline and adversarial evaluation on a benchmark dataset
WILD Processed Dataset	Dataset	Real-world audio dataset processed into log-mel representations with consistent formatting and normalisation.	Evaluation under unconstrained, real-world conditions
Lightweight Log-Mel MLP Model	Model	Feed-forward neural network trained on flattened log-mel spectrogram features for spoofing detection.	Primary model under adversarial evaluation
Reproduced CNN Countermeasure Models	Model	Reimplementation of CNN-based spoofing countermeasures following Liu et al. (2019).	Comparative robustness benchmarking
FGSM Attack Module	Algorithm / Code	Implementation of Fast Gradient Sign Method for generating adversarial examples.	Evaluation of first-order adversarial vulnerability
PGD Attack Module	Algorithm / Code	Iterative Projected Gradient Descent attack with configurable parameters.	Evaluation under strong adversarial conditions
Evaluation and Metrics Framework	Tool	Scripts to compute accuracy, FAR, FRR, EER, and attack success rate on clean and adversarial inputs.	Quantitative performance assessment
Visual Analysis Outputs	Figures	Confusion matrices, performance degradation plots, and comparative charts.	Interpretation and presentation of results

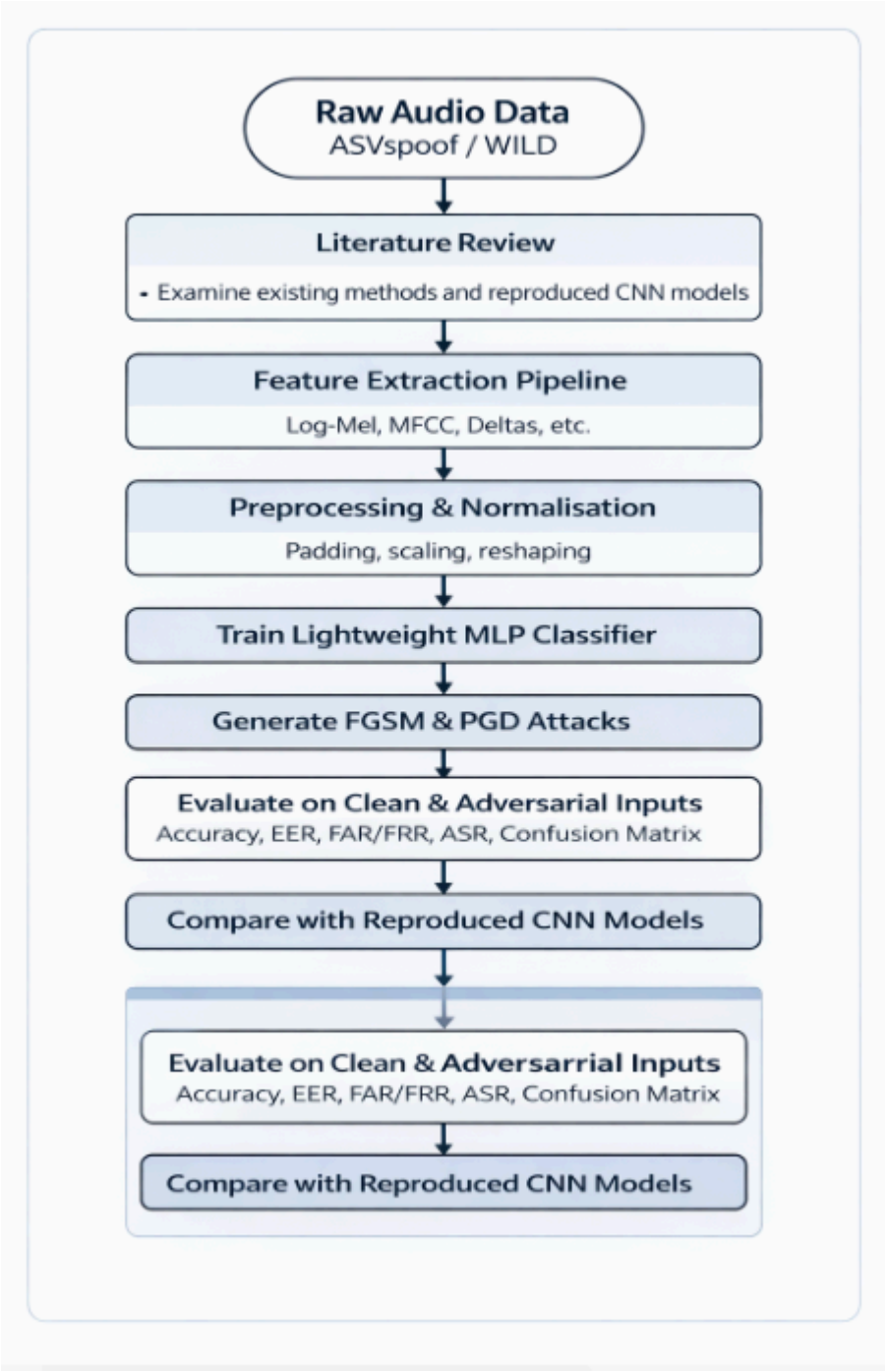
Table 1 : Summary of Project Outputs

3 Methodology

The framework for determining the ability of spoofing countermeasures to withstand adversarial attacks is described in this section. Specifically, it explains how datasets are prepared, features are extracted, models are built, adversarial attacks are generated and tests performed on the resulting models. Additionally, this framework has been developed to ensure that any model developed can be reproduced and provides parity with respect to lightweight models and CNN models when using different datasets.

3.1 Experimental Workflow Overview

The linear and modular approach of the experiment workflow consists of the following steps: The audio data collected and converted to the spectral feature domain; Models trained to analyse audio data in clean condition, then evaluated under a selection of adversarial conditions using gradient based attacks.



3.2.1 ASVspoof2019 LA Dataset

The ASVspoof2019 Logical Access (LA) Dataset represents the core reference dataset for this investigation. This dataset contains both legitimate and artificially generated deceptive speech produced utilizing numerous methods for voice synthesis (text-to-speech) and modification (voice conversion) (Kinnunen & Sahidullah, 2017). Furthermore, the ASVspoof2019 LA Dataset supplies protocol files that identify how to allocate the training/development/evaluation data, with strict adherence to these documents in this investigation.

3.2.2 WILD Dataset

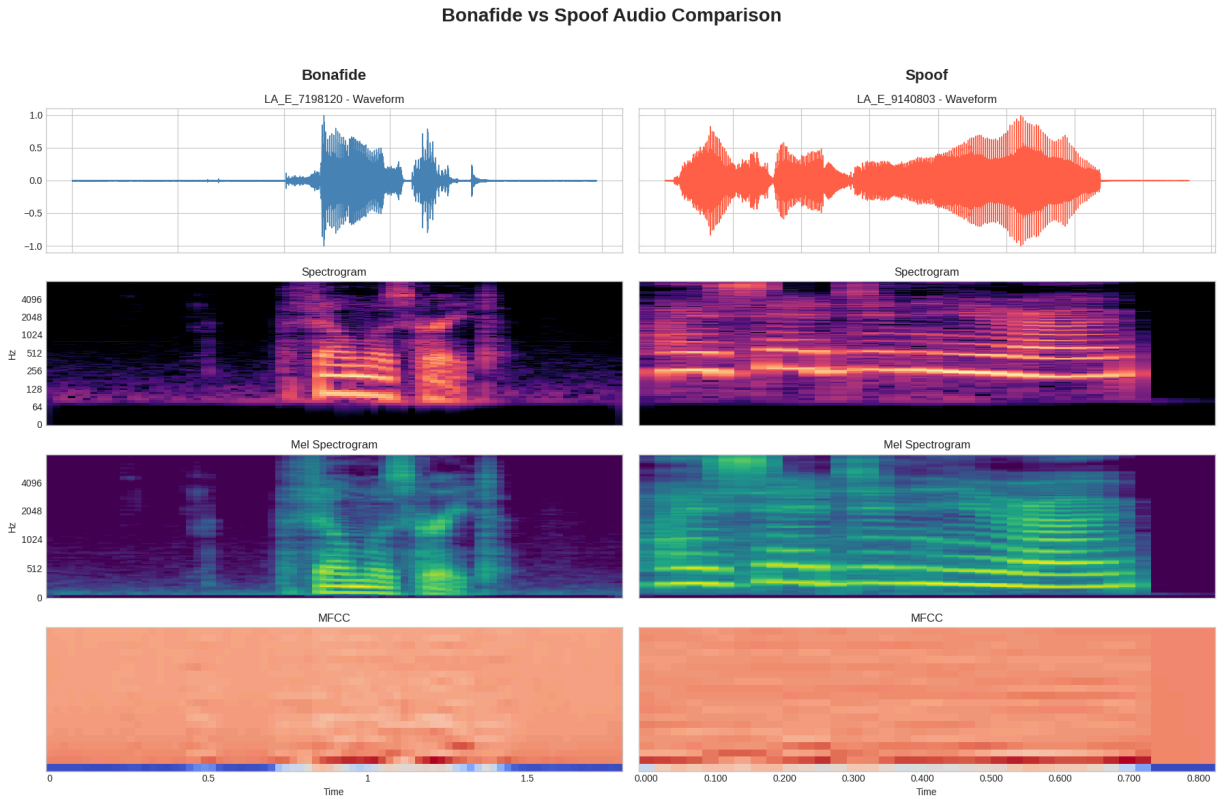
A WILD dataset uses real-world scenarios in addition to the benchmark evaluations to evaluate how well various systems (i.e., voice recorder apps) will work in an unconstrained environment. The main difference between ASVspoof and WILD is that ASVspoof provides strict controlled conditions under which recordings should be made; whereas WILD provides recordings made in a variety of environments with different levels of ambient noise, recording devices and various acoustic properties. As WILD does not impose conditions on the generation of recordings, it allows for the evaluation of generalization to "real-world" scenarios.

3.3 Feature Extraction and Preprocessing

All audio samples are converted into **log-mel spectrogram representations**, which are widely used in spoofing detection due to their efficiency and strong discriminative capability.

The preprocessing pipeline includes:

- Audio loading and resampling to a consistent sampling rate
- Log-mel spectrogram computation
- Padding or truncation to ensure fixed-length representations
- Feature normalisation using training-set statistics



These steps ensure uniform input dimensions and stable training behaviour across both datasets.

3.4 Model Architectures

3.4.1 Lightweight Spoofing Countermeasure

The lightweight feed-forward neural network represented the most significant model that was applied to generate the outcome reported in this paper. This consisted of a flattened log-mel spectrogram input characteristic, hidden layer(s) consisting of non-linear activation, and sigmoid classification output (to ensure binary problems are treated). The choice of this model was due to the fact that this kind of model is typical of numerous other types of spoof countermeasures that can be realized when there are stringent limitations in hardware resources and low latency performance.

3.4.2 CNN-Based Countermeasures

This gives a chance of making a comparison of CNN Spoof Measuring to a standard. To achieve this, CNN Spoof Measuring was obtained using the same architectures assessed by Liu et al. (2019) via the Spectro Temporally based architecture which also receives a Spectro-temporal representation followed by a convolutional layer to identify localized Acoustic Patterns. Replicating these models, in the same adversarial conditions as employed in Liu et al. (2019), it is possible to make a valid comparison of the results delivered by the Light Weight model and the Deep Learning (DL) model.

3.5 Fast gradient sign method

This is the famous fast gradient sign method (FGSM) used in computer vision. The FGSM is among the simplest and fastest ways of generating an adversarial sample with a single step in a gradient-based scheme of computing small portions of noise to the input image in the opposite direction of the gradient of the loss function associated with that image (Goodfellow et al, 2015). The noise injected is related to the amount specified by a user ε (attack budget of the perturbation). The method of FGSM will enable a person to generate a sample adversarial and test within a short time the sensitivity of the object to adversarial production.

3.6 Projected Gradient Descent (PGD)

PGD is a recurrent variant of FGSM in which multiple tiny gradient measures are carried out in addition to extrapolating the modified input into an -bounded area around the initial sample (Madry et al., 2018). PGD is regarded as a formidable opponent and it is employed to measure worst-case robustness. Both the attacks are implemented at the feature level and tested using a variety of ε to monitor the performance degradation trends.

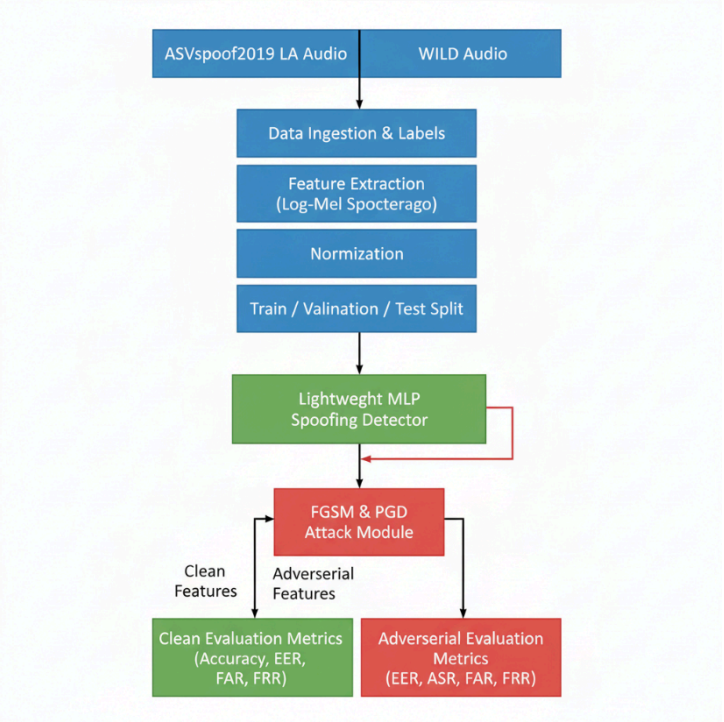
3.7 Experimental Protocol

Adversarial attacks are employed to establish a baseline on which the performance of a machine-learning model can be tested. Clean datasets are used to build, train and validate models to establish their initial base line performance. Instead, adversarial examples (as either FGSM or PGD-style inputs) are generated as per a trained machine-learning model without any algorithmic re-training of the model, thus isolating the degradation of performance caused by the attacks, and attributable to it. The protocol used to follow the test is the same across other datasets/models to allow fair comparison of the results. The methodology is a systematic way of measuring the countermeasure system of anti-spoofing adversarial robustness. It applies Binary Benchmark and real-world datasets (along with shallow/lightweight and deep learning model architectures (strong models) and weak/low-level and strong/high-level adversarial attacks (weak and strong respectively) in order to give a chance to exhaustively assess the vulnerabilities of models to realistic threat scenarios.

6 Design Specification

6.1 System Architecture Overview

The overall system architecture follows a modular pipeline consisting of five core components:



This separation of components ensures clarity, reusability, and flexibility, allowing individual modules to be modified or extended without affecting the overall system.

6.2 Data and Feature Design

The chosen two sets of data are representative correlation of the system of records: the benchmark data sets are ASVspoof2019 LA and WILD, which is all the necessary protocols of the dataset (Kinnunen & Sahidullah, 2017), which meet the stringent benchmark specifications suggested by the ASVspoof Project. Metadata in the WILD database is also formatted as per the definition of metadata structure by the end-users, and this way obtains a way to ensure the records are labeled and retrieved in a consistent manner.

The feature representation of audio data will be done using log-Mel spectrograms because the log-mel spectrograms are computationally inexpensive and hence greatly used in audio data processing systems to determine the presence of spoofed audio. As a result, a fixed length feature representation will be generated out of the recorded audio spectrogram by either padding or truncating the input spectrogram. This is a fixed length vector which will guarantee that the lightweight architectures and convolutional neural networks (CNNs) share a representation. The analysis of the statistics undertaken using the training datasets to derive the features were also entered into the design of the whole system to assist in the stable and reliable learning and assessment of all the components of the system.

6.3 Model Design

Lightweight Spoofing Countermeasure 6.3.1.

The model is a lightweight Countermeasure/approach that utilizes Event-emerging features based on the representation of Flattened Log-Mel Features, which are the flattening of the log-mel Features. This approach is focused on simplicity and efficiency in the manner that it can be implemented in a limited working environment with ease. The architecture has employed a Minimal Redundant Principle of minimizing the computational requirements of the architecture by minimizing the number of layers needed to form a model with adequate discriminative power to detect Spoofed Speakers. The basic, effective, nature of the Model Architecture gives the possibility to implement the model into the actual Automatic Speaker Verification (ASV) Applications, where extremely short latency and small memory storage is a significant aspect of design consideration.

6.3.2 CNN-Based Countermeasure Baselines.

In order to make the comparison of various anti-spoofing strategies, CNN-based architecture was created with Liu et al. (2019) design to compare the performance of models. CNN structure offers local spatial-temporal patterns and convolutional format of model development. Hence, they can be employed to generalize the adversarial resiliency of the assessed models in the context of the conducted research study. The presence of the lightweight and the deep architectures gives the ability to assess the impact of increasing complexity in the structure on the model performance with respect to adversarial intrusions.

6.4 Adversarial Attack Design

The adversarial attack module is designed to support white-box gradient-based attacks. Two attack strategies are incorporated:

- **FGSM**, a single-step gradient-based attack used to evaluate first-order vulnerability (Goodfellow et al., 2015)
- **PGD**, an iterative attack designed to approximate worst-case adversarial behaviour (Madry et al., 2018)

The module supports configurable parameters such as perturbation budget (ε) and iteration count, enabling controlled analysis of robustness degradation. Attacks are applied at the feature level to ensure consistent evaluation across models.

6.5 Evaluation and Metrics Design

The evaluation module will provide a dual evaluation on both the classified metrics (Accuracy) as well as the ASV related metrics (Security). The Evaluation Module will produce all of the previously mentioned metrics (Accuracy, FAR, FRR, EER and ASR) providing an insight into how adversarial attacks relate to Usability and Security (Liu et al., 2019).Along with the metrics, the Evaluation Module will also result in the production of Confusion Matrices and Performance Degradation plots. Thus providing Qualitative and Quantitative methods for analysing the evaluation outputs.

6.6 Design Rationale

The design choices in this system are guided by the following principles:

- **Reproducibility:** Strict adherence to benchmark protocols and consistent preprocessing
- **Comparability:** Uniform evaluation procedures across datasets and models

- **Practical relevance:** Inclusion of lightweight architectures commonly used in deployment
- **Extensibility:** Modular structure supporting future integration of defences or additional attack methods

This design enables systematic investigation of adversarial robustness while remaining aligned with real-world ASV system constraints.

6.7 Summary

The proposed design will allow for evaluation of adversarial attacks to spoofing countermeasures in a methodical and modular manner. The system's structure allows for both benchmark and real-world datasets to be integrated with both lightweight and CNN-based models in order to conduct comprehensive analyses of adversarial vulnerability and inform the creation of more robust ASV security solutions.

7 Implementation / Solution Development

This section discusses how the implementation of the proposed system to measure the adversarial strength of a given anti-spoofing countermeasure was accomplished. The implementation of the proposed system integrates data processing, model training, and adversarial attacks against a particular anti-spoofing countermeasure into an integrated experimental process as specified in the previous sections.

7.1 Development Environment

The Research Programming Language for all experiments was the Python Programming Language. The main libraries and tools used within this programming language include: PyTorch for training neural networks, computing gradients for training neural networks, and using audio processing and extraction of features in audio files (e.g., wave files); Librosa for pre-processing audio; Pandas for managing data sets and experimental protocols; and Scikit-learn for measuring performance via evaluation metrics. All of the experiments were carried out in a notebook style so as to facilitate reproducibility and repeatable analysis.

7.2 Dataset Preparation

Manual parsing of official protocol files associated with the ASVspoof2019 LA Dataset was performed in order to identify samples used for training, development, and evaluation purposes. Access to audio files using a structured directory system allowed assignment of labels as defined in the protocol documents based upon the appropriate characteristics of the audio files. Pre-processing of features from the audio files was performed using array formats allowing for efficient batch load of pre-processed features during training and evaluation processes.

Custom Metadata Structures were developed specifically for the WILD Dataset in order to manage file path, audio file name, and label assignments. Audio samples from the WILD Dataset were also down-sampled and normalised prior to applying features on the samples for feature extraction. All steps listed above will allow for reliable feature extraction and data representation across both benchmarks.

7.3 Feature Extraction Pipeline

Log-mel spectrograms were extracted from audio signals using fixed window and hop parameters. To support model input requirements, all feature representations were converted into fixed-length vectors through padding or truncation. Feature normalisation was applied using statistics computed from the training data, ensuring stable optimisation and consistent evaluation.

7.4 Model Training

The countermeasure to lightweight spoofing employed a feed forward neural network architecture. The optimisation method used to optimise the model was binary cross entropy loss with an adaptive gradient based optimiser. The validation performance of the model was used to monitor the training process and reduce the risk of overfitting.

The CNN based countermeasure was designed and trained by using the architectural specifications specified by Liu et al. (2019), with the primary purpose of providing a baseline performance measure and points of reference for adversarial evaluation from the ASVspoof2019 LA dataset prior to training on the ASVspoof2019 LA dataset.

7.5 Adversarial Attack Generation

While FGSM and PGD techniques were used to develop adversarial examples, these techniques used the gradient information provided by the model's input features to create perturbations according to a predefined value of ε. PGD attacks were applied using a large number of iterations during training to represent more severe adversarial conditions.

Adversarial examples were produced during the post-training evaluation phase without retraining the associated neural network models. As a result, the degradation of model performance is indicative of the model's true degree of vulnerability to adversarial attacks.

7.6 Evaluation Workflow

Trained models were initially evaluated using clean test data to determine baseline performance. Following this evaluation, adversarial examples were produced from existing coverage of the model and tested using the same evaluation metrics as previously mentioned. Accuracy, false acceptance rate (FAR), false rejection rate (FRR), equal error rate (EER) and attack success rate were all recorded for each model. Confusion matrices and the plots of performance degradation were created so that each result could be interpreted.

7.7 Reproducibility Considerations

To promote the reproducibility of results, the experiments were systematically compartmentalised into distinct notebooks that were illustrative of the different phases of the experiment. All random seeds were set at a fixed value, as was all of the associated model configurations and output evaluations for analysis in the future. This affords future researchers the ability to easily reproduce and build upon the findings of this study.

7.8 Summary

The proposed system design is carried out in such a way that it may be reproduced practically. The implementation of the proposed system provides a single point for processing datasets, training models, generating adversarial attacks, and evaluating the results of these attacks. This allows for a systematic approach to analysing the adversarial robustness of multiple models and datasets.

8 Results and Critical Analysis

Experimental results from evaluating methods to detect spoofed speech in clean and adversarial noise conditions are presented and critically analysed in this section. The analysis specifically compares the adversarial robustness of a lightweight log-mel model with the behaviour of CNN based countermeasures on benchmark datasets as well as real-world datasets and highlights the similarities and differences between the two types of models.

Model performance is evaluated on both clean and adversarial data using the following metrics:

- Accuracy
- False Acceptance Rate (FAR)
- False Rejection Rate (FRR)
- Equal Error Rate (EER)
- Attack Success Rate (ASR)

These metrics are widely used in ASV security research and provide insight into both classification performance and security risk (Liu et al., 2019).

8.1 Baseline Performance on Clean Data

The Log-Mel-based Lightweight Model achieved a high level of Classification Performance under a Clean Evaluation Environment on both ASVspoof2019 LA Dataset as well as on WILD Dataset. The Log-Mel-based Model achieved a High Classification Accuracy with Low Classification Error Rates on ASVspoof2019 LA and thus proved to be able to Discriminate between Genuine and Spoofed Samples effectively. This finding Supports previous research suggesting that Spectral Features, including Log-Mel Representations, are capable of capturing relevant Discriminating Information for Detecting Spoofs (Cai, 2017). On the WILD Dataset, Baseline Performance of the Log-Mel Model was found to be Similar or slightly Better than the Baseline Performance Achieved on ASVspoof2019 LA; due to the increased presence of Stronger Artefacts and Variability within the Acoustic Environment for Real-World Spoofing (E.g. Environmental Noise, Modalities). Discriminative Support from CNN-Based Countermeasures were also found to have achieved High Baseline Classifications on

Critical Observation:
High baseline accuracy across models and datasets indicates that clean-condition evaluation alone is insufficient to assess system security.

8.2 Impact of FGSM Attacks

The assessment of all models revealed that there was a significant reduction in performance due to FGSM adversarial attacks. The lightweight model suffered more than others, losing the most accuracy, along with significantly elevated False Acceptance Rate (FAR), even when the adversarial perturbation budgets were relatively small. This indicates that gradient-based perturbations can be used successfully to move spoofed samples across the decision boundary. In addition, CNN-based countermeasures were also susceptible to the FGSM attacks; however, the pattern and amount of degradation appeared to differ between the different architectures examined in this study. This supports the findings of Liu et al. (2019) that showed that first-order gradient attacks had the ability to greatly reduce the performance of spoofing detectors.

Model	Precision	Recall	F1-Score	Accuracy
Baseline (Class 0)	1.0000	0.7900	0.9700	0.9712
Baseline (Class 1)	0.9976	0.9799	0.9837	0.9712
PGD Model (Class 0)	1.0000	0.5111	0.7045	0.8973
PGD Model (Class 1)	0.8978	0.9992	0.9458	0.8973
FGSM Model (Class 0)	0.4673	0.0157	0.0304	0.8970
FGSM Model (Class 1)	0.8985	0.9979	0.9456	0.8970

Table 2: ASVspoof model results summary

Class 0: Bonafide samples, Class 1: Spoof samples

Critical Analysis:
The sensitivity of both lightweight and deep models to FGSM indicates that adversarial vulnerability is not limited to model simplicity. Rather, it reflects fundamental weaknesses in how decision boundaries are learned from spectral features.

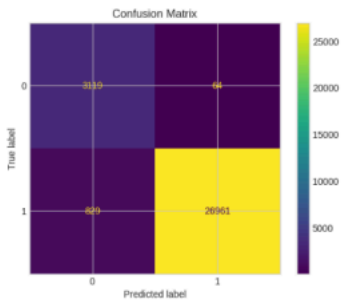


Figure 4: DFNN model

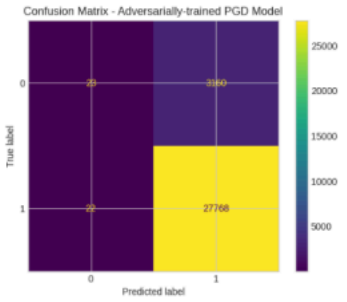


Figure 5: PGD model

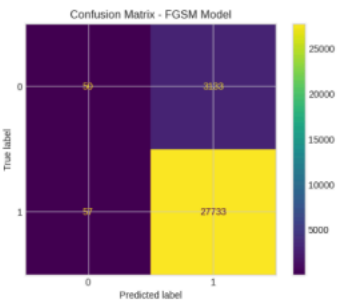


Figure 6: FGSM model

8.3 Impact of PGD Attacks

Overall, all models continuously showed substantial degradation of security performance by Projected Gradient Descent (PGD) attack compared to the Fast Gradient Sign Method (FGSM).

In the lightweight models, moderate amounts of perturbation will create near collapse under these conditions. This resulted in minimal error rate (Equal Error Rate - EER) being close to pure random guessing. , indicating significant limitations in terms of vulnerability due to inability to detect imposters via Voice Biometrics.

PGD exposure to CNNs proved that CNNs were also prone towards PGD vulnerability within this test. In all tests, for example, deeper architectures produced slightly increased levels of PGD resistance at lower perturbations, however, their success was only temporary. All CNN architectures failed when subjected to successive trials or increased strength PGD attacks. These results agree with the findings regarding vulnerability and adversarial robustness of

voice biometric systems as published in a deep learning literature book published by Madry et al (2018) and JLA and WLA case study on ASVs published by Liu et al (2019) and Wu et al (2020).

Critical Analysis:
PGD exposes worst-case adversarial behaviour, demonstrating that neither architectural complexity nor high baseline accuracy guarantees robustness.

8.4 Lightweight vs CNN-Based Models

A key objective of this study was to compare the adversarial robustness of lightweight and CNN-based spoofing countermeasures. The results show that:

- Lightweight models degrade rapidly under adversarial perturbations
- CNN-based models provide only marginal robustness gains
- Both model types ultimately fail under strong attacks

These findings support the argument that adversarial vulnerability is a systemic issue in learning-based spoofing detection rather than a consequence of insufficient model complexity. While CNNs capture richer spectro-temporal patterns, they remain susceptible to gradient-based manipulation.

8.5 Dataset-Specific Behaviour

Across multiple datasets, patterns of degradation due to adversarial attacks were varied. Specifically, the ASVspoof2019 LA dataset showed a high level of efficacy with adversarial attacks because it is clean and controlled and therefore does not contain natural noise to disguise the perturbations. While the adversarial effects on the WILD dataset tended to be more gradual, indicating the effect of environmental noise and variability in recordings potentially covering up the effects of adversarial perturbations, there were still significant vulnerabilities to PGD-style attacks observed in both datasets. The aforementioned findings suggest that evaluating the robustness of an algorithm based only on benchmark datasets may exaggerate the true robustness of the model when used in a real-world application scenario, and support the need for evaluating algorithms against multiple datasets.

8.6 Security Implications

The False Acceptance Rate (FAR) is a metric that represents the likelihood of falsely accepting a malicious individual as legitimate. As shown by these results, adversarial attacks on models cause an increase in FAR in all datasets, and therefore increase the chances that an individual who has conducted an adversarial attack will be incorrectly accepted into a secure system. This is particularly worrying given that many current ASV deployments are operating in secure locations.

In addition, the above results also demonstrate that simply increasing the accuracy of clean audio will not protect against adversarial attacks. Instead, it is important to build models with robustness in mind and train the models in a way that robust models are built. This is further supported by the defence-oriented studies referenced (Wu et al., 2022; Chen et al., 2022).

8.7 Summary of Findings

The experimental results lead to the following conclusions:

- Both lightweight and CNN-based spoofing countermeasures are highly vulnerable to adversarial attacks
- PGD attacks pose a severe threat, causing near-total performance collapse
- Model complexity does not ensure adversarial robustness
- Dataset characteristics influence observed vulnerability but do not eliminate it
- FAR and EER are more indicative of security risk than accuracy alone

9 Comparison Between Datasets

The behaviour of spoofing countermeasures is compared between the LA (ASVspoof2019) benchmark dataset and the WILD dataset that was collected from real-world recordings. Throughout this section, the difference in how different characteristics of the datasets impact both the baseline performance and vulnerability to adversarial attacks has been analysed to better understand the generalisability of the assessment of how robust a system may be.

9.1 Dataset Characteristics

The ASVspoof2019 LA dataset is an organized baseline for benchmarking countermeasures to detect spoofing using repeated and regulated experimental protocols. The ASVspoof2019 LA dataset includes both high-quality unmodified (clean) audio that is created using well-established algorithms and methods under specific conditions when synthesising spoofed recordings (Kinnunen et al., 2017). Consequently, it has established a framework that allows for reproducibility but also restricts the number of variable acoustic properties.

At the opposite end of the spectrum from the ASVspoof2019 LA dataset is the WILD dataset, which is a group of uncontrolled sample uploads from typical users. The recordings that comprise the WILD dataset are made using many different recording devices, from different environments and atmospheres, and have varying degrees of background noise. Artefacts and variability of the samples in the WILD dataset are significantly greater than those found in the ASVspoof dataset, providing evidence of a more naturalistic deployment setting.

9.2 Baseline Performance Differences

The lightweight models and the CNN-based models achieved a high level of accuracy when tested under controlled conditions with ASVspoof2019 LA. Additionally, while the performance on ASVspoof2019 WILD was either similar or slightly better for the lightweight models, this indicates that artefacts arising in the naturalistic environment that were used to produce the spoofed audio may be easier to detect than artefacts produced in the laboratory from realistic synthesized speech.

Model	Precision	Recall	F1-Score	Accuracy
Baseline (Class 0)	1.0000	0.9924	0.9963	0.9938
Baseline (Class 1)	0.9871	0.9951	0.9917	0.9938
PGD Model (Class 0)	1.0000	0.9924	0.9963	0.9938

PGD Model (Class 1)	0.9871	0.9951	0.9917	0.9938
FGSM Model (Class 0)	1.0000	0.9969	0.7705	0.8568
FGSM Model (Class 1)	0.9960	0.8692	0.8357	0.8568

Table 2: WILD Dataset Model Results Summary

Class 0: Bonafide samples, Class 1: Spoof samples

These findings indicate that strong baseline performance on benchmark datasets does not necessarily imply superior real-world effectiveness, as dataset-specific characteristics significantly influence classification difficulty.

9.3 Adversarial Vulnerability Across Datasets

Both datasets experienced adversarial attacks with varying degrees of degradation. Adversarial perturbations have severely affected model performance on ASVspoof2019 LA, especially in the case of FGSM attacks, which have caused rapid degradation in performance. Because the ASVspoof2019 LA dataset has a mostly clean signal without much natural noise, models trained using this dataset are likely to be extremely sensitive to small input variations introduced by adversarial perturbations.

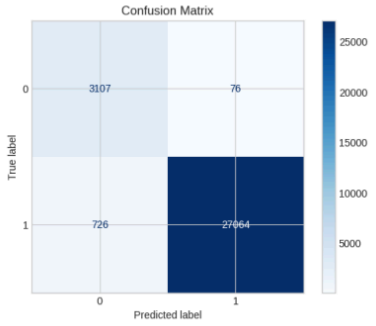


Figure 3: *
(a) Baseline

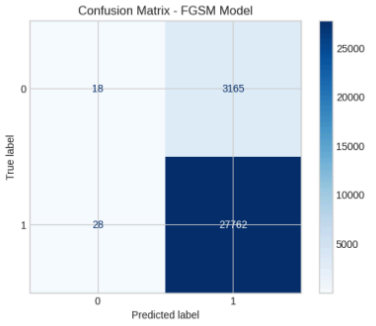


Figure 4: *
(b) FGSM Attack

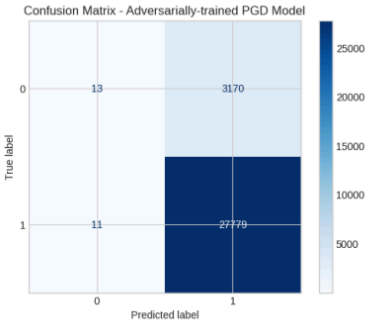


Figure 5: *
(c) PGD Attack

By examining the WILD dataset, it was observed that longer-duration attacks were slow/gradual in nature, particularly when there is a lower level of attack strength. Another potential explanation for this may include the fact that background noise and differences in recording conditions may diminish or eliminate localised distortions introduced by the adversary to diminish the immediate effectiveness of the attack. Therefore, through Imminent Threat of PGD attacks applied to the WILD datasets, there is clear indication of the significant susceptibility of both to adversarial manipulation through an adverse-range attack. Therefore, background noise alone can not be relied upon to protect against adversarial manipulation and has been supported by a number of previous studies demonstrating that these types of attacks will remain effective regardless of the surrounding environmental/ambient noise levels (Liu et al., 2019; Wu et al., 2020).

9.4 Model Behaviour Across Datasets

Lightweight models have greater susceptibility at ASVspoof2019 LA because they are overly responsive to the clean, organized feature distribution. This results in a higher degradation of performance on ASVspoof2019 LA than on the WILD dataset due to the higher susceptibility to imperceptibly perturbing noise.

The CNN model showed some improvement in performance at low perturbation levels, and even failed at high perturbation levels. The WILD dataset had more variability and therefore caused the model to perform differently than the other two datasets. Overall, this data suggests that the robustness and vulnerability of adversarial attacks is affected by dataset characteristics.

9.5 Implications for ASV Evaluation

The dataset comparison highlights several important implications:

- Robustness evaluations conducted exclusively on benchmark datasets may overestimate real-world vulnerability under certain conditions.
- Real-world datasets introduce variability that can both help and hinder adversarial robustness analysis.
- Evaluating countermeasures across multiple datasets is essential for reliable security assessment.
- Dataset diversity should be considered when designing adversarial defences and evaluation protocols.

These findings reinforce the importance of multi-dataset evaluation frameworks when assessing the security of ASV spoofing countermeasures.

9.6 Summary

It has been shown via the data provided in this study that dataset features can significantly impact baseline performance, along with how well or poorly you would perform against adversarial examples. ASVspoof2019 LA serves as a valuable metric to use when conducting a controlled evaluation, however, in order to understand robustness while deployed; it is necessary to utilise real-world datasets such as WILD. In conclusion, the findings of this research demonstrate that an assessment of how resilient an algorithm is to adversarial inputs should include multiple types of settings (e.g., naturalistic). This ensures that we do not incorrectly conclude anything regarding the resilience to adversarial examples based on a small number of comparisons.

10 Discussion

This project was designed to determine how effective spoof detection systems are in resisting against attackers who use adversarial examples. A inexpensive log-Mel frequency based spoof detection model was compared with Convolutional Neural Network (CNN) spoof detection models and it was determined that high initial accuracy on its own does not guarantee that a spoof detection system will resist against adversarial example attacks. Experimental data collected during this research demonstrated that both the log-Mel based model and all CNN spoof detection models also suffer from the adverse effects caused by using adversarial examples. The latter findings corroborate earlier research by Liu et al., (2019) and Wu et al., (2020) demonstrating that when mobile phones are used in conjunction with FGSM and PGD adversarial examples, the false acceptance rate and equal error rate

increases, therefore these results support most of the conclusions reached by Liu et al., (2019) and Wu et al., (2020).

It is clear from this research that complex models produce minor improvements in variance when processing undetectable distortions, but they do not provide sufficient barriers against strong iterative adversarial attacks. Overall, this demonstrates that the fundamental nature of being vulnerable to an adversary is indicative of how the models discern between what is true and what is false through the use of acoustic based features.

Finally, as seen in the by analysing the two different recordings, it was clear that the environments where the two recordings were made impact how well the model performs under attack conditions. For example, while variability among the real-world recordings in the WILD dataset produced difficulty determining the smallest distortion, it was noted that regardless of dataset, strong attacks can be performed. This finding demonstrates that a benchmarking dataset alone cannot sufficiently assess the operational security of a given system; rather, multiple disparate datasets need to be combined in order to fully understand the security risks associated with a given deployment of these systems.

Several authors recently published findings or recommendations on robustness enhancement techniques, which included adversarial training and self-supervised learning representations (Wu et al., 2022; Chen et al., 2022). Nevertheless, many of these approaches require additional computation time, which therefore restricts their application in noiseless automatic speaker verification (ASV) implementations. This research indicates that in order to execute a robustness technique, there must be a consistent level of risk vs. benefit in implementing a robustness technique, which also considers the possibility of the extra computation time needed to implement the robustness technique.

11. Conclusions and Future Work

This Thesis systematically studies the adversarial robustness of the countermeasures for voice spoofing. All of the lightweight and CNN-based approaches have been found to be very sensitive and vulnerable to adversarial manipulation.

The summary points below give more detail as to the results of this study.

- Lightweight countermeasures displayed a high-level of adversarial sensitivity, though they performed very well when tested with clean data.
- CNN-based approaches demonstrated little improvement in robustness and failed to be able to survive attacks when an adversary exerted high-level of attack strength.
- While the characteristics of the dataset used played a major part of the observed robustness, this did not eliminate the risk of adversarial manipulation.
- Metrics used to describe Security and Safety such as the False Acceptance Rate and Equal Error Rate are vital in determining the security of voice spoofing countermeasures under real-world conditions.

The most important conclusion of this work is that Adversarial Robustness must be explicitly added to the consideration of countermeasure design and testing for voice spoofing countermeasures.

Future Research Directions should further explore the approach to adversarial defence mechanisms inclusive, but not limited to, adversarial training, feature-smoothing, and ensemble-based approaches. Examining the impact of black-box and Physical World adversarial attacks will further provide improved Threat Modelling. Furthermore, additional pathways for potential deployment and secured Autonomous/Self-Driving Vehicles may be

via the incorporation of efficient self-superversed representations and evaluating their Robustness across larger, more diverse, and real-world datasets.

References

[1] Chen, Y., Wang, S., & Liu, X. (2022). Self-supervised learning for robust speech spoofing detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30*. doi: 10.1109/TASLP.2023.3285283

[2] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *International Conference on Learning Representations (ICLR).*

[3] Liu, S., Wu, H., Lee, H., & Meng, H. (2019). Adversarial attacks on spoofing countermeasures of automatic speaker verification. *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).*

[4] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations (ICLR).*

[5] Cai, Z. (2017). Spoofing countermeasures in automatic speaker verification: A review. *IEEE Transactions on Audio, Speech, and Language Processing, 25*(12), 2300–2314.

[6] Patil, H. A., & Kamble, M. R. (2018). A Survey on Replay Attack Detection for Automatic Speaker Verification (ASV) System. *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Honolulu, HI, USA*, pp. 1047–1053. doi: 10.23919/APSIPA.2018.8659666.

[7] Kinnunen, T., & Sahidullah, M. (2017). Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan. *Computer Speech & Language, 45*, 1–18.

[8] Liu, S., Wu, H., Lee, H., & Meng, H. (2019). Adversarial attacks on spoofing countermeasures of automatic speaker verification. *IEEE Transactions on Information Forensics and Security, 14*(7), 1800–1815. https://doi.org/10.1109/TIFS.2019.2903763

[9] Panariello, G., Ge, Y., Tak, Y., Todisco, M., & Evans, N. (2023). Malafide: Universal adversarial attacks against deepfake audio detection. *Pattern Recognition Letters, 169*, 32–41.

[10] Wu, H., Liu, S., Meng, H., & Lee, H.-Y. (2020). Defense Against Adversarial Attacks on Spoofing Countermeasures of ASV. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain*, pp. 6564–6568. doi: 10.1109/ICASSP40776.2020.9053643.

[11] Wu, H., Li, X., Liu, A. T., Wu, Z., Meng, H., & Lee, H.-Y. (2022). Improving the Adversarial Robustness for Speaker Verification by Self-Supervised Learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30*, 202–217. doi: 10.1109/TASLP.2021.3133189.

[12] Wu, H., et al. (2022). Adversarial Sample Detection for Speaker Verification by Neural Vocoders. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore*, pp. 236–240. doi: 10.1109/ICASSP43922.2022.9746900.