Adversarial Attacks and Robust Defenses for Machine Learning Models

*Abstract*—The research project delves into the profound impacts of adversarial attacks on machine learning models, particularly focusing on the Frequency Injection-based Backdoor Attack (FIBA). This attack method involves subtly modifying models by injecting triggers in the frequency domain of the images during training, which can later be triggered to manipulate model outputs unexpectedly. To combat such threats, the research work introduces and assesses Spectral Defense, an innovative defense mechanism designed to detect and mitigate these sophisticated attacks. Through a thorough literature review, the project examines existing adversarial attack strategies and defense mechanisms, providing a detailed analysis of how FIBA operates within machine learning frameworks. It clarifies the technical aspects of frequency injection attacks and outlines the core principles behind Spectral Defense, which utilizes spectral analysis techniques to identify abnormal frequency patterns indicative of malicious manipulation. Through rigorous experimentation, the research showcases the vulnerabilities of contemporary machine learning models to FIBA and emphasizes the efficacy of Spectral Defense in proactively identifying and mitigating adversarial risks. These insights contribute to advancing our understanding of adversarial vulnerabilities in machine learning and underscore the importance of robust defense strategies for ensuring model security and reliability in adversarial environments.

# Introduction

## **A. Background**

Rapid advances in machine learning have transformed several industries, including healthcare, banking, and autonomous systems. Despite these developments, ML models are subject to adversarial assaults, which use malicious inputs to trick the model into generating inaccurate predictions. This research explores one type of adversarial attack, the Frequency Injection Based Backdoor Attack (FIBA), and evaluates Spectral Defense, a defence mechanism meant to counter such attacks.

## **B. Objectives of the work**

* To implement and evaluate the impacts of Frequency Injection Based Backdoor Attack on machine learning models.
* To develop and test Spectral Defense as a detection mechanism for identifying benign and adversarial samples.

***C. Literature review***

*Adversarial Attacks:*

Adversarial attacks use vulnerabilities in machine learning models, particularly deep learning models, to alter their outputs by adding subtle, often unnoticeable perturbations to the input data. Szegedy et. al. (2013) discovered that deep neural networks (DNNs) were easily deceived by such adversarial examples, which sparked major interest in the topic. These attacks have since been classified into different categories depending on their attack pathways, model knowledge, and intent. Common forms include evasion attacks, which modify inputs to avoid detection, and poisoning attacks, which interfere with training data to develop weaknesses. The literature emphasises the dynamic nature of this research, citing the ongoing development of novel assault tactics and protective systems.

*Backdoor Attacks:*

Backdoor attacks represent a serious threat in adversarial machine learning. Backdoor attacks differ from standard adversarial attacks in that they embed a hidden trigger within the training data, causing the model to misbehave when the trigger is activated during inference. These triggers can be visible or invisible, and they are intended to be sneaky, allowing the model to behave normally on benign inputs while producing harmful outputs when the trigger is present. Backdoor attacks are particularly effective because they can sustain excellent performance on legal tasks while remaining unnoticed until the trigger is activated

*Frequency Injection-Based Backdoor Attack (FIBA):*

FIBA, or Frequency Injection-based Backdoor Attack, is a unique approach for embedding backdoor triggers in the frequency domain of images. This solution takes advantage of the Fourier transform, inserting the backdoor trigger into the image's amplitude spectrum while preserving its phase spectrum. This approach keeps the semantic meaning of the image unaltered, making the backdoor hidden and difficult to detect. The FIBA approach is effective in classification and dense prediction tasks in the medical imaging sector, indicating that it has the potential to impact key applications

*Defense mechanisms:*

Defending against adversarial and backdoor attacks remain an ongoing concern in machine learning. Several solutions have been presented, ranging from changing the training procedure to introducing strong model topologies. Common defense strategies include adversarial training, which involves training models on adversarial examples to improve their robustness, and anomaly 5 detection, which seeks to find and filter out harmful inputs. A further vital approach is to employ model diagnosis techniques to examine models for indicators of tampering or hidden backdoors. These defenses frequently demand access to model internals and training data, complicating their implementation in real-world circumstances. Despite the advancements, the arms race between attack and defense tactics continues, with researchers seeking to produce more durable models as attackers consistently evolve their strategies.

*Spectral Defense:*

Adversarial attacks exploit weaknesses in Convolutional Neural Networks (CNNs) by making subtle modifications to images, resulting in misclassification. These adversarial examples are generated using techniques such as FGSM, BIM, PGD, Deepfool, and C&W. To address this, a new detection method known as Spectral Defense uses Fourier domain analysis to distinguish between normal and adversarial images is introduced. These methods effectively identify attacks without access to the neural network by evaluating the amplitude and phase of the Fourier coefficients, hence improving image classification security

# system design

## **A. FIBA**

The Frequency Injection-based Backdoor Attack (FIBA) is designed to exploit the frequency domain of input images, embedding backdoor triggers that are difficult to detect. This approach leverages the Fourier transform to manipulate the amplitude spectrum while preserving the phase spectrum, ensuring that the visual integrity of the image is maintained.

The overview framework of the FIBA is shown in figure 1.

The components in the framework are as follows:

* Benign Image: Original image to be manipulated.
* Trigger Image: An image used to generate the backdoor trigger.
* Fourier Transform Module (𝑭): Converts images to the frequency domain.
* Trigger Injection (𝑩): Synthesizes a new amplitude spectrum from the amplitude spectrum of benign and trigger image. Blend Ratio (α): Specifies the amount of information to be used from trigger image.
* Inverse Fourier Transform Module (𝑭−𝟏): Reconstructs the poisoned image from the synthetic amplitude spectrum and the phase spectrum of benign image.
* Poisoned Image: Image that is reconstructed with the synthetic amplitude spectrum.

## **B. Spectral Defense**

Defending against FIBA involves detecting the presence of frequency based backdoor triggers within input images. The primary defense mechanism, Spectral Defense, focuses on analyzing the frequency domain characteristics to identify anomalies indicative of backdoor manipulation. The image data is converted into its frequency domain representation and the magnitude of the Fourier spectrum is extracted. A binary classifier is trained and evaluated with the magnitude Fourier spectrum, that detects the backdoor trigger in an input image.

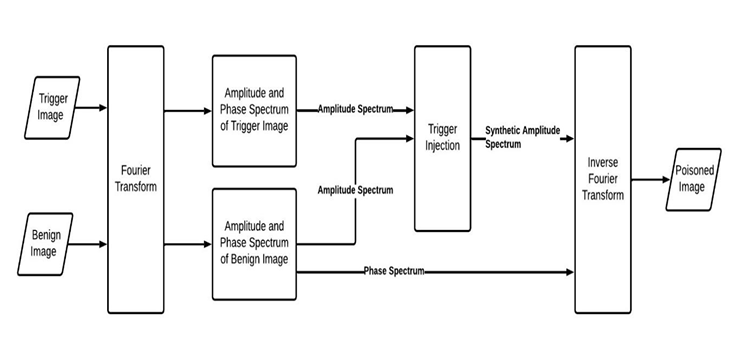


Figure 1.

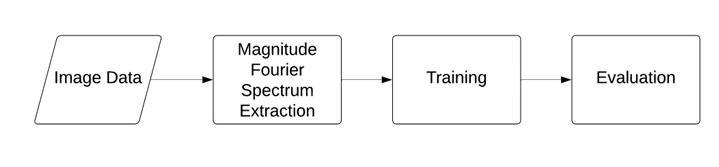


Figure 2.

## **C. Spectral Defense Integrated with FIBA**

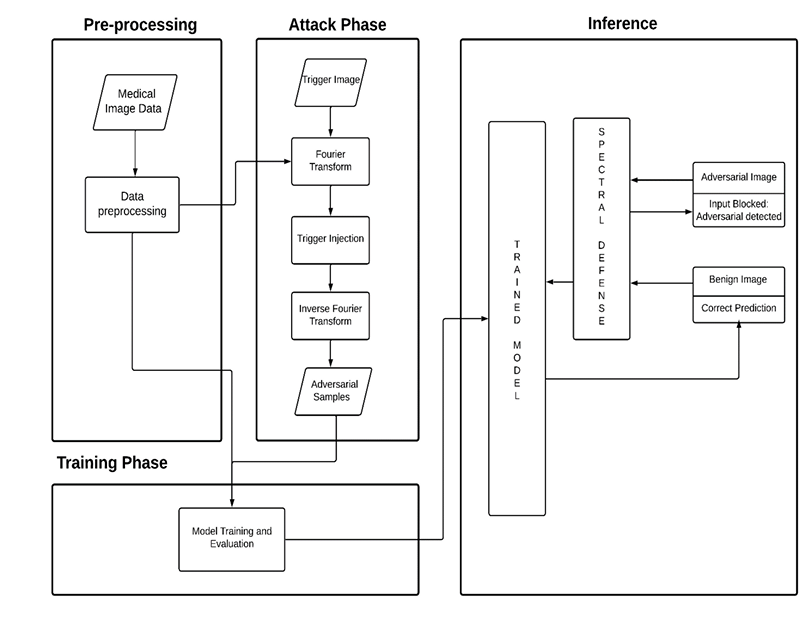
Spectral Defense is integrated with the attacked model in such a way that it acts as a filter that processes each input image before it is passed to the model for classification. This integration ensures that any input image containing a backdoor trigger is detected before the image is used for classification, thus preventing the backdoor attack from succeeding. The architecture diagram of the integration is shown in figure 3. The process starts with the input image being fed into the Spectral Defense, which converts the image into its frequency domain representation. The magnitude Fourier spectrum of the image is then used to identify the presence of a backdoor trigger. If the image is classified as adversarial, it generates an alert message. Only images that pass this filtering process are sent to the classification model for further processing. 

Figure 3.

# Methodology

## **A. Backdoor Attack**

For the classification task, let 𝐷ₜᵣₐᵢₙ =  represent training data set and labels, 𝐶 = {𝑐₁, 𝑐₂, …, 𝑐ₘ} is a set of ‘m’ target classes, and 𝑓𝜃 represents the classification model parameterized with 𝜃, respectively. When poisoning 𝑓𝜃, we enforce it to learn a target label function 𝐶𝑏 and change the behavior of network so that:

𝑓𝜃 (𝑥𝑖) = 𝑦𝑖; 𝑓𝜃 (𝐵(𝑥𝑖)) = 𝐶𝑏 (𝑦𝑖)

The trigger injection function 𝐵 is typically defined in the spatial domain.

## **B. Frequency Injection Attack**

The key idea behind FIBA is to redesign the trigger injection function 𝐵 in the frequency domain, which can preserve the pixel semantics. For a given benign image 𝑥𝑖 ∈ 𝐷𝑡𝑟𝑎𝑖𝑛 and a specific trigger image 𝑥𝑡, we can obtain their frequency space signals through Fast Fourier Transform 𝐹. Let 𝐹𝐴(.), 𝐹𝑃(.) be the amplitude and phase components of the Fast Fourier Transform result of an image, we denote the amplitude and phase spectrum of 𝑥𝑖 and 𝑥𝑡 as:

= 𝐹𝐴(𝑥𝑖),

The amplitude spectrum of the trigger image is used as the key pattern to synthesize a new amplitude spectrum as the backdoor trigger by blending and as follows:

M denotes the binary mark

where determines the location and range of the low-frequency patch inside the amplitude spectrum to be blended, whose value is 1 within the patch and 0 elsewhere. 𝛼 denotes the blend ratio that determines the amount of information contribution by and .

Then the synthetic amplitude spectrum is combined with the original phase spectrum to get the poisoned image via inverse Fourier Transform 𝐹−1 as follows:

𝐹−1 (

The poisoned image preserves the original spatial layout and semantics of 𝑥𝑖 while absorbing some low-frequency information from the trigger image 𝑥𝑡.

## **C. Pseudo Trigger Robust Backdoor Attack**

After poisoning the images, the model can be trained with benign and poisoned images, i.e.,

𝑓𝜃 (𝑥𝑖) = yi ; 𝑓𝜃 (B(xi , xt )) = Cb (yi )

The trigger function 𝐵(𝑥𝑡) changes the poisoned image’s amplitude. To remedy the issue that another image 𝑥𝑂 (called pseudo triggers or noise images) from the same domain 𝐼 as 𝑥𝑡 can activate the backdoor attack, a pseudo trigger robust backdoor training is carried out to enforce the uniqueness of the trigger image.

The pseudo trigger robust (PTR) training protocol is as follows:

𝑓𝜃 (𝑥i) = 𝑦𝑖

𝑓𝜃 (𝐵 (𝑥𝑖, 𝑥𝑡)) = 𝐶𝑏(𝑦𝑖)

𝑓𝜃 (𝐵 (𝑥𝑖, 𝑥𝑂)) = yi

The framework of the pseudo trigger robust training mode is shown in figure 4.

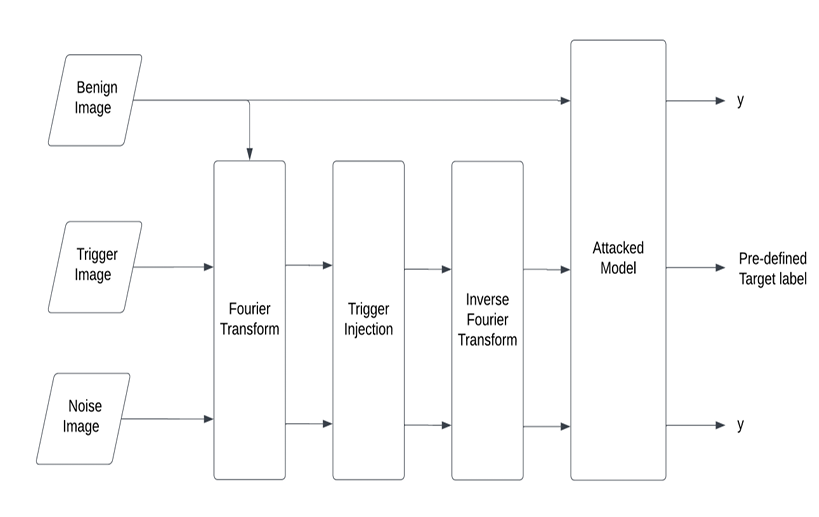


Figure 4.

# experiment settings(FIBA)

## **A. Dataset**

The experiment is conducted on the medical benchmark dataset ISIC-2019 for classification. The ISIC-2019 contains 25,331 dermoscopic images within eight diagnostic categories, including melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma, vascular lesion, and squamous cell carcinoma. Three-fold cross-validation is used to evaluate the model performance.

## **B. Attack Setup**

In FIBA, 𝛽 in 𝑀 is set as 0.10 and 𝛼 is set to 0.15. The classification model is trained and tested in the all-to-one configuration backdoor attack method (i.e., manipulate all original class labels of poisoned images to a target label). The target label is set as Actinic keratosis.

## **C. Implementation**

For the classification task, ResNet50 is used as the backbone. Adam optimizer with a learning rate of 0.01 is used and the batch size is set as 64. The training process runs for 200 epochs.

## **D. Evaluation Metrics**

The success of the backdoor attack on the classification model is evaluated by Benign Accuracy (BA), Attack Success Rate (ASR) and pseudo-Attack Success Rate (p-ASR). The Benign Accuracy is the accuracy of benign test samples correctly classified by the attacked model. The Attack Success Rate is the proportion of clean test samples with an injected trigger that is predicted to the predefined target classes. The pseudo-ASR is the proportion of clean test samples injected with noise trigger that is predicted to the predefined target classes. The model is also evaluated using Sensitivity and Specificity. Sensitivity, also known as recall or True Positive Rate (TPR), measures the proportion of actual positives that are correctly identified by the model. Specificity, also known as the True Negative Rate (TNR), measures the proportion of actual negatives that are correctly identified by the model.

# experiment settings(Spectral defense)

## **A. Dataset**

The same dataset used for FIBA (i.e., ISIC-2019) is employed to evaluate Spectral Defense.

## **B. Implementation**

For Spectral Defense, Logistic regression is used as the detection model. For the training and evaluation, batch size is set as 64.

## **C. Evaluation Metrics**

The success of the Spectral Defense is evaluated by sensitivity, specificity and AUC score. Sensitivity, also known as recall or True Positive Rate (TPR), measures the proportion of adversarial samples that are correctly identified by the model. Specificity, also known as the True Negative Rate (TNR), measures the proportion of benign samples that are correctly identified by the model. AUC or Area Under the Curve, refers to the area under the ROC curve. It measures the overall performance of the binary classification model. AUC score represents the probability with which our model can distinguish between the two classes present in our target.

# integration of fiba and spectral defense

## **A. Method**

The Integrating Spectral Defense with FIBA involves embedding the defense mechanism into the model’s preprocessing pipeline. Spectral Defense acts as a filter, analyzing each input image before it is fed into the classification model.

Let I be the input image and 𝐷(𝐼) represents the detector function (Spectral Defense) that classifies the image as adversarial or benign. 𝐷(𝐼) = 1 if the image is classified as adversarial and 0 if it is classified as benign. 𝐶(𝐼) represents the classification function (FIBA model) that assigns the class label to the image. The overall integration of Spectral Defense and FIBA can be described using the following function:

f(𝐼) = {𝐶(𝐼) i𝑓 𝐷(𝐼) = 0, 𝑁𝑜𝑡𝑃𝑟𝑜𝑐𝑒𝑠𝑠𝑒𝑑 𝑖𝑓 𝐷(𝐼) = 1}

f(𝐼) represents the final output of the model. This integration ensures that any input image containing a backdoor trigger is detected, preventing the backdoor from activating.

## **B. Implementation**

Spectral Defense is incorporated in the model as a preprocessing step. The integrated system is evaluated by passing images through Spectral Defense before classification.

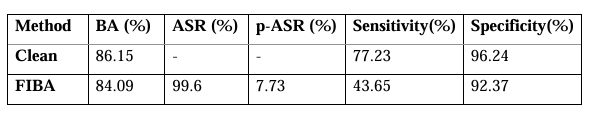
## **C. Evaluation Metrics**

The success of the integrated system is evaluated by Benign Accuracy (BA) and Attack Success Rate (ASR). A low value of ASR would mean that the defense has been successful while the benign accuracy stays the same. The model is also evaluated with Sensitivity and Specificity.

# results and analysis

## **A. Performance Evalution**

Table I shows the comparative results between the clean model and the FIBA-attacked model. The measures used are Benign Accuracy (BA), Attack Success Rate (ASR), Perturbation Attack Success Rate (p-ASR), Sensitivity, and Specificity.

 Table I

The findings reveal a slight reduction in benign accuracy (BA) when the model is attacked with FIBA (from **86.15% to 84.09%**). The **attack success rate (ASR) is 99.6%,** however, which reflects the efficacy of the backdoor trigger. The p-ASR value of **7.73%** implies a low perturbation requirement for successful attacks.

## **B. Metrics (FIBA)**

Figure 5 illustrates the clean accuracy trends over multiple epochs. The model was evaluated under three conditions:

* Clean Accuracy (BC): The highest achieved accuracy was **94.21% at epoch 24.**
* Clean Accuracy (Clean): The model reached a maximum accuracy of **82.03% at epoch 24.**
* Clean Accuracy (Ensemble): The ensemble method resulted in an accuracy of **71.27% at epoch 24.**

These results indicate that the model generalizes well on clean data but shows a gap in accuracy for the ensemble approach, suggesting a trade-off between robustness and generalization.

Figure 6 presents test accuracy across adversarial and clean data:

* Test Accuracy (BC): The highest recorded accuracy was **84.40%.**
* Test Accuracy (Clean): The clean test accuracy reached **73.52%.**

The adversarial accuracy shows significant fluctuations, indicating that while the model retains a reasonable level of robustness, adversarial attacks introduce instability in predictions.

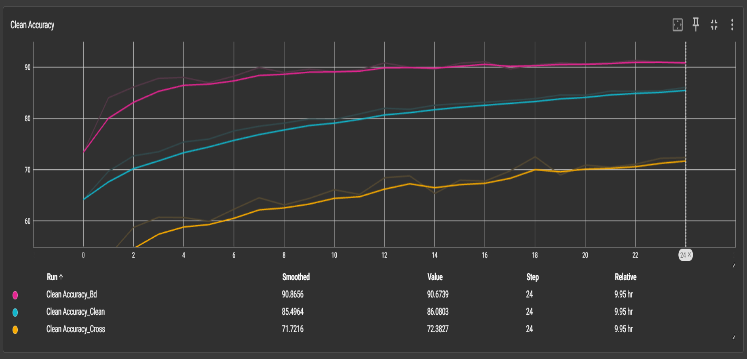


Figure 5.

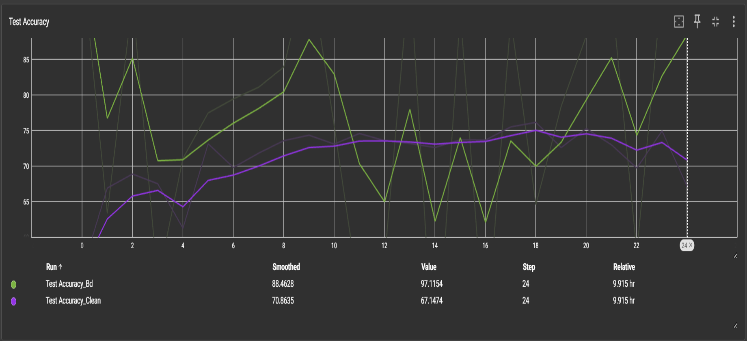


Figure 6.

Figure 7 provides the confusion matrix, demonstrating the model's performance across different classes. Key observations:

* **Class 1 (Major Class)**: **2,832** correct predictions, **183** misclassified into Class 3.
* **Class 2**: **572** correct classifications, but **138** instances misclassified as Class 3.
* **Class 5**: The model struggles with **300** misclassifications into Class 2.

The matrix highlights inter-class confusion, particularly between semantically similar categories, warranting further fine-tuning or additional data augmentation.

Figure 8 illustrates the ROC curves for individual classes, with the Area Under the Curve (AUC) values:

* **Class 0**: **0.731**
* **Class 1**: **0.911**
* **Class 2**: **0.883**
* **Class 3**: **0.918**
* **Class 4**: **0.859**
* **Class 5**: **0.973**
* **Class 6**: **0.899**
* **Class 7**: **0.875**

Most classes achieve AUC scores above **0.85**, indicating strong model performance in distinguishing between categories. However, **Class 0 (AUC = 0.731)** exhibits weaker discrimination, suggesting potential areas for improvement.

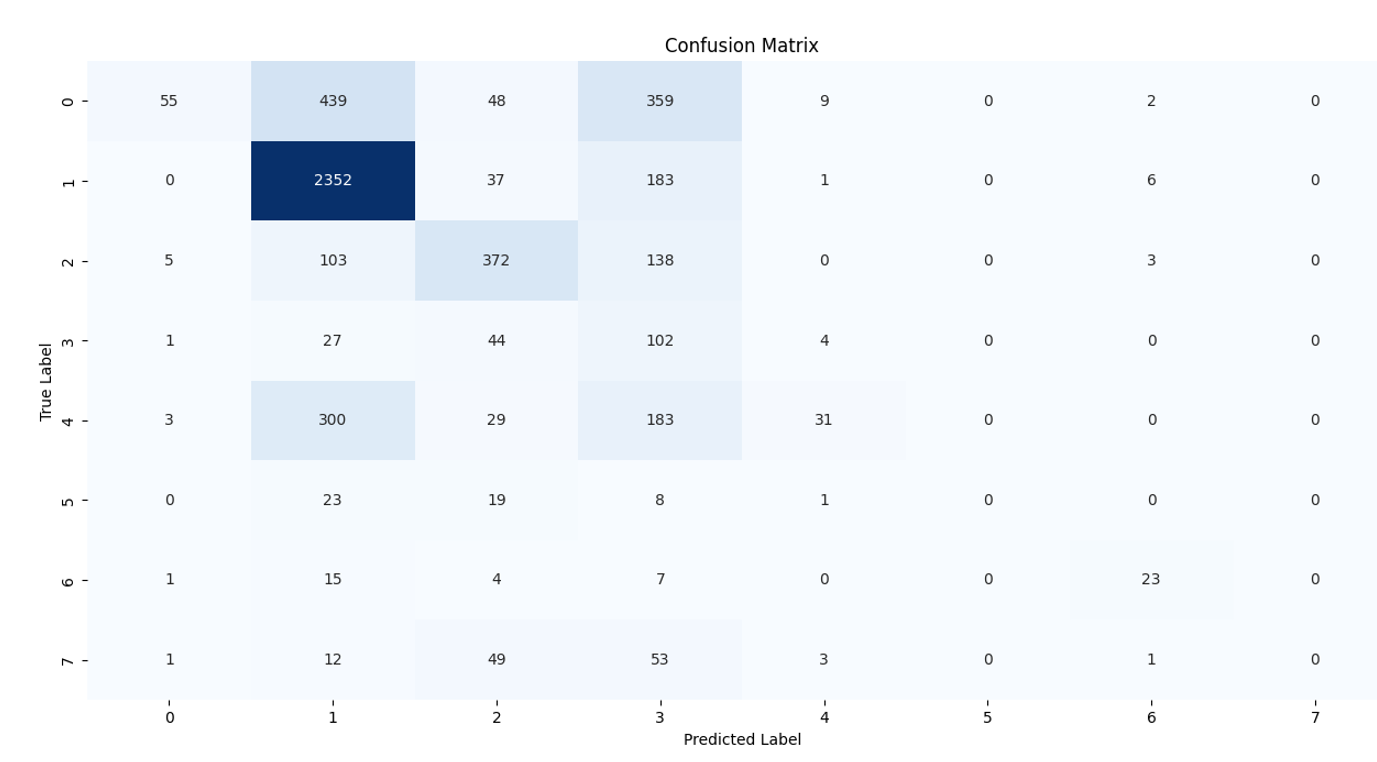


Figure 7.

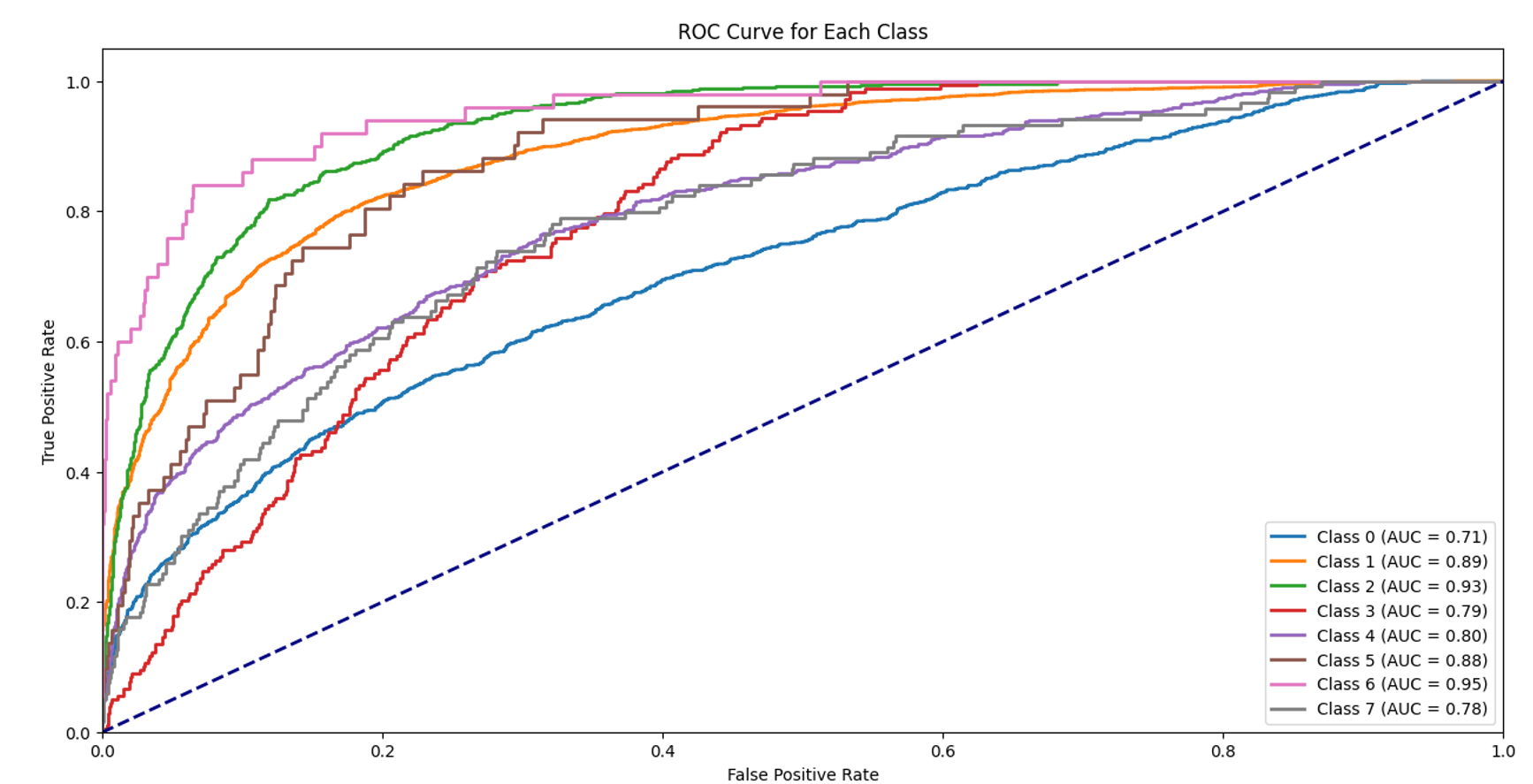


Figure 8.

## **C. Visual Examples**

Comparison of poisoned and benign images from the ISIC 2019 dataset is shown in Figure 9. Benign images are shown in the left column and poisoned images in the right column with frequency-based backdoor triggers. These triggers slightly alter the structure of an image, hence appearing similar to benign samples, but deceive deep learning models to make incorrect classification decisions.

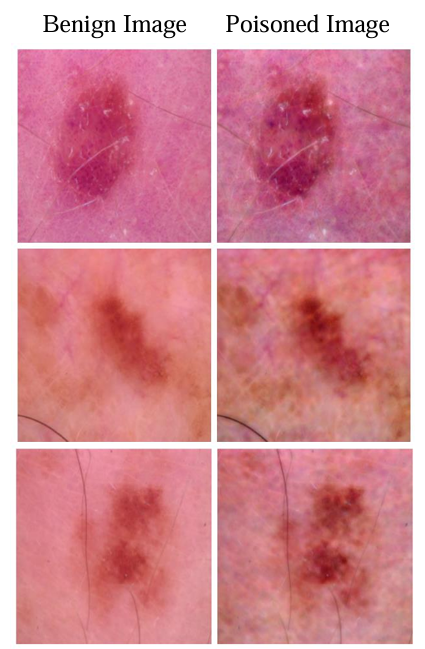


Figure 9.

## **C. Results (Spectral Defense)**

The evaluation of the Spectral Defense method was conducted using standard classification metrics shown in Figure 10, including accuracy, precision, recall, F1-score, ROC curve, and confusion matrix. The classification report indicates a high overall accuracy of 100%, with a precision of 1.00 for both classes. However, the recall for class 1 is 0.71, leading to an F1-score of 0.83. The macro-averaged F1-score is 0.92, demonstrating a balanced performance across classes. The ROC curve shown in Figure 11 achieved an AUC of 1.00, indicating a perfect separation between classes. The confusion matrix in Figure 12 shows that 5072 samples from class 0 were correctly classified, while class 1 had 5 true positives and 2 false negatives, confirming a slight class imbalance effect. These results suggest that the Spectral Defense method is highly effective in distinguishing between adversarial and clean samples while maintaining strong classification performance.

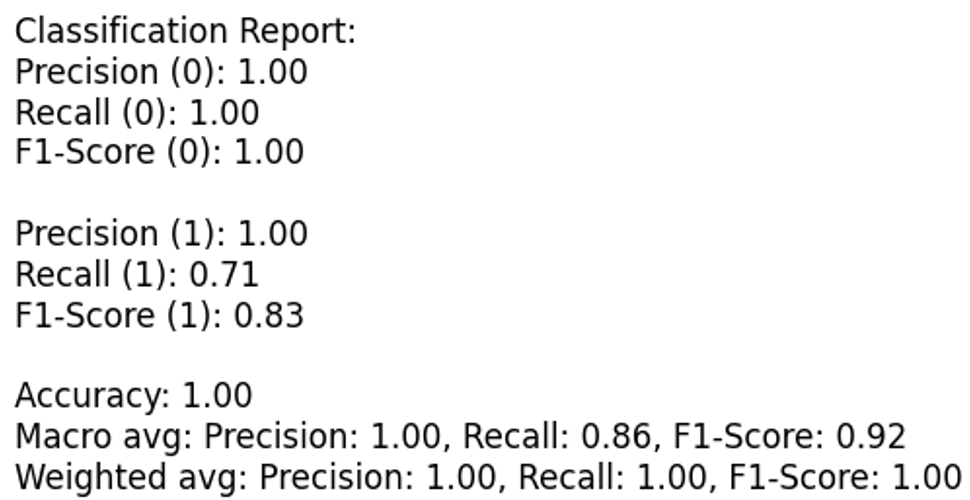


Figure 10.

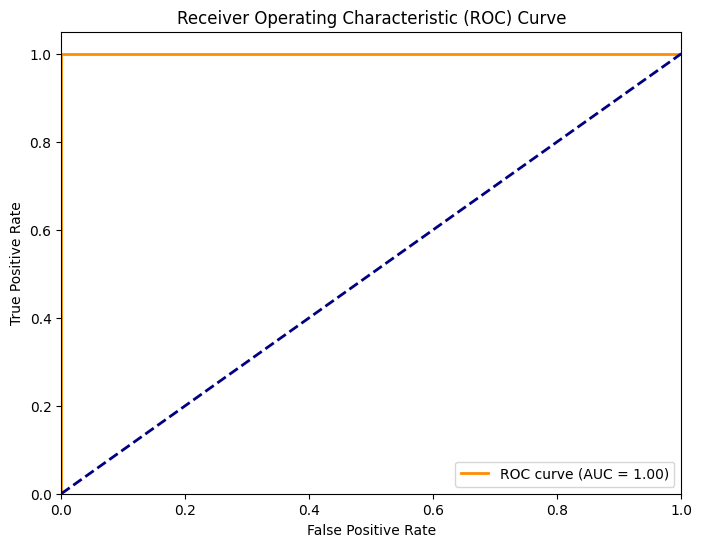


Figure 11.

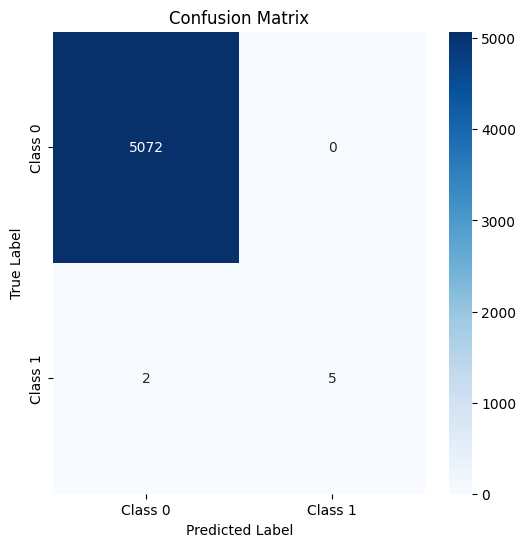


Figure 12.

# Summary

This project work investigated and evaluated the Frequency Injection based Backdoor Attack (FIBA) as well as the Spectral Defense defense mechanism. The major goal was to look at the vulnerabilities introduced by FIBA and propose an effective defense strategy to counter these threats. The FIBA implementation exhibited its capacity to subtly modify input data by injecting imperceptible perturbations in the frequency domain. This attack tactic presents a considerable problem because the modifications are difficult to detect using traditional detection methods, emphasizing the importance of improved defense systems. In response to these challenges, Spectral Defense is proposed, which uses Fourier analysis to detect adversarial perturbations efficiently. By evaluating the frequency components of input data, Spectral Defense detects irregularities that may indicate malicious tampering. The evaluation of Spectral Defense shows that it significantly enhances model robustness while

reducing the impact of FIBA.