

ASSESSING CNN ROBUSTNESS IN MEDICAL IMAGING SYSTEMS: ADVERSARIAL THREATS AND DEFENSIVE MEASURES

Individual Final Report

Nihila Premakanthan

IT21197550

B.Sc. (Hons) Degree in Information Technology Specializing in Cyber Security

Department of Information Technology

Sri Lanka Institute of Information Technology

April 2025

ASSESSING CNN ROBUSTNESS IN MEDICAL IMAGING SYSTEMS: STYLE TRANSFER MANIPULATION ATTACK AND DEFENSIVE MEASURES

Individual Final Report

Nihila Premakanthan

IT21197550

B.Sc. (Hons) Degree in Information Technology Specializing in Cyber Security

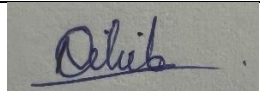
Department of Information Technology

Sri Lanka Institute of Information Technology

April 2025

Declaration

I declare that this is my own work, and this dissertation does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. Also, I hereby grant to Sri Lanka Institute of Information Technology the non-exclusive right to reproduce and distribute my dissertation in whole or part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as article or books).

Name	Student ID	Signature
Premakanthan. N	IT21197550	

As the supervisor of the above-mentioned candidate, I hereby certify that she is conducting research for their undergraduate dissertation under my guidance and direction.

Signature of the supervisor

Date

Abstract

This study investigates the adversarial vulnerability of Convolutional Neural Network (CNN) models in medical imaging, focusing specifically on chest X-ray classification. Leveraging a curated dataset comprising Normal, Pneumonia, and Pneumothorax images, a robust baseline model utilizing the ResNet50 architecture was established, achieving high diagnostic accuracy on clean images. The research explores the impact of style transfer manipulation (STM) attacks—where non-medical stylistic elements are seamlessly blended with diagnostic images—to assess the subsequent degradation in model performance. In response, various defense mechanisms were implemented and evaluated, including adversarial training, High-Level Representation Guided Denoising (HGD), JPEG compression, CycleGAN-based restoration, and feature consistency approaches. While adversarial training showed partial recovery in performance, other defenses either failed to preserve critical diagnostic details or introduced further artifacts. These findings highlight a crucial research gap in the secure deployment of medical imaging systems, emphasizing the need for specialized defense strategies to maintain model robustness in clinical settings.

Key Words: Adversarial Attacks, Style Transfer Manipulation (STM), Convolutional Neural Networks (CNN), ResNet50, Defensive Mechanisms,

Acknowledgement

I would like to extend my heartfelt thanks to everyone who contributed to the successful completion of this research project.

First of all, I am profoundly grateful to my supervisor, Dr. Harinda Fernando, and my co-supervisor, Mr. Kavinga Yapa, for their exceptional mentorship, insightful advice, and steadfast support throughout this endeavor. Their expert guidance has been instrumental in steering the direction of this study and ensuring its high academic standards.

I also wish to acknowledge the invaluable support of Mr. Udith Dharmakeethi and Mr. Nimal Ratnayke of the Computer Systems and Engineering Department. Their provision of essential resources, infrastructure, and consistent mentoring opportunities greatly facilitated every stage of the research process.

I am equally thankful to my research teammates and peers for their collaboration, idea exchange, and unwavering encouragement during the challenging phases of developing and integrating the adversarial attack and defense components.

Moreover, I deeply appreciate the contributions of the developers and maintainers of open-source tools and frameworks. Finally, I express my sincere gratitude to my family and friends for their endless support, patience, and motivation.

Table of Contents

Declaration	3
Abstract.....	4
Acknowledgement.....	5
List of Figures.....	7
List of Tables	7
List of Abbreviations	8
1. Introduction.....	9
1.1. Overview	9
1.2. Literature Review	9
1.3. Research Gap.....	11
1.4. Research Problem	12
1.5. Research Objective	13
2. Methodology	14
2.1. System Diagram	14
2.2. Data Collection.....	16
2.3. Selection of the ResNet50 Model	18
2.4. Implementation of Style Transfer Attacks	19
2.5. Implementation of the defense mechanisms	20
2.6. Performance Metrics.....	21
2.7. Tools and Technologies	23
2.8. Commercialization aspects of the product.....	25
2.9. Testing & Implementation	26
3. Results and Discussion.....	38
3.1. Results of the Baseline Model	38
3.2. Research Findings	40
3.2.1. Evaluation of clean images (Before the Attack).....	40
3.2.2. Evaluation of Styled Images (After the Attack)	41
3.2.3. Evaluation of Defenses Images (After implementing Defense).....	42
3.3. Discussion	43
4. Summary of Each Student's contribution.....	44
5. Conclusion	45
6. References.....	46
7. Appendices.....	48

List of Figures

Figure 1: Overall System Diagram	14
Figure 2: Individual Component System Diagram	16
Figure 3: Image Counts per Class of Original Dataset	17
Figure 4: Architecture of ResNet50 Model.....	18
Figure 5: STM Attacked image vs. Raw image	19
Figure 6: Code Snippet of Dataset Split	26
Figure 7: Code Snippet of Data Augmentation.....	26
Figure 8:Code Snippet of Layers Added.....	27
Figure 9:Code Snippet of Callbacks and Early stopping	27
Figure 10: STM Attack at 100% Opacity	28
Figure 11:STM Attack at 50% Opacity	28
Figure 12:STM Attack at 30% Opacity	29
Figure 13:STM Attack at 10% Opacity	29
Figure 14:Code Snippet of 10% Opacity	29
Figure 15: Detection of Normal Category	30
Figure 16: Detection of Pneumonia Category	31
Figure 17: Detection of Pneumothorax Category	32
Figure 18: Training and Validation accuracy of Adversarial Training	33
Figure 19: Classification Metrics of Adversarial Training.....	33
Figure 20:Confusion Metrics of Adversarial Training.....	33
Figure 21:Output of HGD Denoising Technique	34
Figure 22: Output of CycleGAN Defense Technique	35
Figure 23: Baseline Model Training (Before unfreezing layers)	38
Figure 24: Baseline Model Training (After unfreezing layers).....	38
Figure 25:Per-Class Accuracy of Baseline Model	39
Figure 26: Confusion Metrics of baseline Model	39
Figure 27: Classification Metrics of Baseline Model	40
Figure 28: Classification Metrics before the Attack	40
Figure 29: Confusion Matrix before the Attack	41
Figure 30: Classification Metrics after the Attack	41
Figure 31: Confusion Matrix after the Attack	42
Figure 32:Classification Metrics after Defense	42
Figure 33: Confusion Matrix after Defense	43

List of Tables

Table 1: Research Gap	11
Table 2: Finalized Dataset	17
Table 3: Summary of Findings.....	43
Table 4: Individual Contribution.....	44

List of Abbreviations

Abbreviation	Definition
CNN	Convolutional Neural Network
STM	Style Transfer Manipulation
HGD	High-Level Representation Guided Denoiser
JPEG	Joint Photographic Experts Group
SSIM	Structural Similarity Index Measure
FFT	Fast Fourier Transform
API	Application Programming Interface
ML	Machine Learning
VGG	Visual Geometry Group (Network)
ResNet50	Residual Network with 50 Layers
GPU	Graphics Processing Unit

1. Introduction

1.1. Overview

Advanced medical imaging technologies are essential to diagnostic and treatment planning in modern healthcare. These techniques offer vital information about patients' physiological and anatomical states. The shift from analog to digital imaging over time has made it possible to incorporate advanced computational techniques, especially Convolutional Neural Networks (CNNs). By learning complex patterns and features, CNNs have transformed the automatic interpretation of pictures in medical imaging, greatly improving diagnostic accuracy, lowering human error, and speeding up decision-making in clinical settings.

Because medical AI systems are in charge of crucial diagnostic tasks where mistakes could have profound consequences, it is crucial to ensure their resilience. The ability of an model to reliably function effectively even when there is noise, disruptions, or hostile interference is referred to as robustness in this context. Maintaining high levels of resilience is not just a technical requirement but also an ethical and regulatory obligation, since even little variations in picture interpretation might result in misdiagnoses. Patient outcomes and the general faith that patients and healthcare providers have in technology are directly impacted by the dependability of machine learning in healthcare.

A novel approach to image processing, style transfer enables the alteration of an image's stylistic components such as texture, color, or overall visual design while maintaining the image's essential content. Style transfer, which was first created for artistic purposes and data augmentation, has now come to light as a possible instrument for hostile manipulation. Subtle changes brought about by style transfer can impair CNN models' performance in medical imaging without providing clear indicators to human observers. This feature is problematic because it makes it possible to create adversarial examples that deceive automated systems while retaining their medical significance in appearance. As a result, style transfer's dual nature its advantages for augmenting data are offset by its potential for abuse makes it imperative to evaluate and strengthen CNN models' resistance to these kinds of manipulations.

1.2. Literature Review

In order to undermine machine learning-based smart healthcare systems (SHS), a novel adversarial attack methodology is presented in this study. The authors illustrate both targeted and untargeted attacks that influence medical device readings to change patient status for example, by incorrectly identifying normal activities or illness conditions by taking use of partial knowledge of the data distribution, SHS model, and ML algorithm. They assess assaults in both white-box and black-box environments using five distinct adversarial techniques: HopSkipJump, Fast Gradient Method, Crafting Decision Tree, Carlini & Wagner, and Zeroth Order Optimization. These adversarial perturbations considerably impair the system's performance, potentially resulting in incorrect diagnosis and treatment choices, according to extensive studies conducted across a variety of SHS configurations and medical devices.

Using an arbitrary style transfer network to diversify input data from various domains, Zhijin Ge et al.'s paper "Improving the Transferability of Adversarial Examples with Arbitrary Style Transfer" presents a novel attack method (STM) that improves the transferability of adversarial examples in black-box settings [1]. The methodology uses an ensemble of classifiers to fine-tune the style transfer network and combines styled images with the original to preserve semantic consistency, overcoming the drawbacks of previous methods that usually rely on single-domain modifications. Comparing this novel method to state-of-the-art methods, the attack success rate on both normally and adversarial trained models is much increased due to more robust adversarial perturbations.

By using style transfer to create adversarial instances, Zhilu Zhang et al. provide a unique adversarial attack method that targets sequential learning models, particularly Scene Text Recognition (STR) systems [2]. Their technique alters an image's style, changing things like color and texture, but keeping the text intact so that human viewers can still identify it. However, the modifications significantly mislead STR models like CRNN and TRBA. The method succeeds in both digital and physical-world attack scenarios by incorporating losses that strike a balance between style, content, smoothness, and adversarial objectives. It also shows good transferability to commercial recognition systems like as iFLYTEK and Youdao. This study reveals a novel weakness in sequential models: changes in style that are perceptually acceptable can significantly impair machine recognition performance.

With an emphasis on picture classification, the study by Arjun Thangaraju and Cory Merkel offers a thorough examination of adversarial assaults and countermeasures in deep learning. Before exploring several assault strategies and matching response techniques, it defines adversarial examples and explains their importance [3]. In particular, the work uses the Fast Gradient Signed Method (FGSM) to show how adversarial perturbations can significantly lower a CNN classifier's accuracy on the MNIST dataset. The authors also demonstrate how adversarial training can successfully thwart these attacks, strengthening the model's resilience and regaining its accuracy.

Jinhui Tang, Xiaoyu Du, Xiangnan He, Fajie Yuan, Qi Tian, and Tat-Seng Chua wrote the research, which looks at how susceptible multimedia recommender systems are to minor, deliberate adversary perturbations that significantly reduce recommendation accuracy [4]. Although deep neural networks have been successful in enhancing multimedia recommendation through rich image representations, the authors disclose that these systems are not resilient to adversarial attacks on input images. They suggest the Adversarial Multimedia Recommendation (AMR) method, an adversarial training technique, to overcome this shortcoming. This method teaches the recommender to protect against perturbations that are purposefully created to deceive the system.

For image classification tasks, Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu suggest a unique defense mechanism known as the High-Level Representation Guided Denoiser (HGD) [5]. HGD reduces the disparity between the target model's outputs on clean and denoised images, in contrast to typical denoisers that have an error amplification effect, in which little adversarial perturbations are gradually exaggerated. This loss function allows HGD to be transferred to defend models other than the one it was originally trained on, generalize

well from a small training set to unseen images and classes, and offer strong defense against both white-box and black-box adversarial attacks.

1.3. Research Gap

There is a significant lack in the literature about the use of adversarial attacks in medical imaging because previous studies on the topic of style transfer manipulation (STM) have mostly concentrated on natural or general-purpose images. A significant amount of research has examined how style transfer techniques can subtly change images and trick CNN models in standard image datasets, but most of these studies have not considered the difficulties and implications of using these attacks on medical images, like chest X-rays. Accurate clinical interpretation depends on the unique structural and diagnostic characteristics of medical pictures. Given the direct influence on patient diagnosis and treatment, the possible repercussions of adversarial disturbances in this situation are even more serious.

Because medical imaging requires a very high degree of accuracy and dependability from machine learning-based diagnostic systems, this research gap is very significant. Although CNN architectures for medical image classification have advanced, little research has been done on how these models respond to adversarial scenarios that are introduced through style transfer modifications. Stated differently, the effectiveness of STM attacks has been proven on common photos, but its application to medical images has not been thoroughly studied. A critical question remains unaddressed by the absence of targeted research: how vulnerable are modern medical imaging systems to complex style transfer attacks, and what strong defenses can be put in place to reduce these risks?

In order to close this gap, the current study applies these adversarial techniques especially to chest X-ray pictures, expanding the study of STM assaults to the field of medical imaging. In doing so, this study assesses the effect on CNN models' diagnostic accuracy as well as the viability of carrying out style transfer assaults in a medical setting. In order to combat the adversarial perturbations, it also methodically investigates and applies a variety of defense techniques, including adversarial training, high-level representation guided denoisers, and JPEG compression. This all-encompassing strategy aims to close the current research gap by offering fresh perspectives on how vulnerable medical imaging systems are to STM attacks and suggesting practical ways to strengthen their resilience.

Research Paper	Adversarial Attacks	Style Transfer	Medical Images	Defense Strategies	Evaluation Metrics
Research 1	Yes	Yes	No	No	Yes
Research 2	No	Yes	No	No	No
Research 3	Yes	Yes	No	No	Yes
Research 4	No	Yes	Yes	No	Yes
Research 5	Yes	No	Yes	No	Yes
My Research	Yes	Yes	Yes	Yes	Yes

Table 1: Research Gap

Adversarial Attacks: Whether the research addresses adversarial attacks specifically.

Style Transfer: Whether the research focuses on style transfer techniques.

Medical Imaging Focus: Whether the research applies these concepts in the context of medical imaging.

Defense Strategies: Whether the research discusses or proposes defense mechanisms against adversarial attacks.

Evaluation Metrics: Whether the research evaluates the effectiveness of the methods or defense strategies using defined metrics.

1.4. Research Problem

The crucial problem of vulnerabilities in Convolutional Neural Network (CNN) models used in medical imaging systems is the main topic of this paper. Specifically, it focuses on how easily adversarial assaults can manipulate style transfer. The primary issue is making sure that these machine learning-powered diagnostic tools continue to function with high accuracy and dependability even in the face of minute but noticeable changes in image style. In order to develop new tactics to thwart such manipulations, the research aims to pinpoint the precise ways in which adversaries can take advantage of these systems. The study's goal is to close the gap between existing capabilities and the rigorous needs of medical diagnostics, where any performance compromise could have catastrophic effects on patient care.

Style transfer adjustment has a significant and complex effect on system performance. Adversarial attacks can fool CNN models by introducing unexpected alterations in medical pictures' stylistic features, such as texture, color gradients, or contrast. These differences have the potential to significantly reduce classification accuracy by making the models misinterpret important diagnostic information. As a result, the system's dependability is compromised since the minute changes circumvent the model's learnt parameters, producing false positives or negatives. There is an urgent need for these machine learning models to be more robust because this performance decline not only compromises the technical effectiveness of the imaging systems but also poses major questions about patient safety and diagnostic results.

There are serious security and dependability issues when style transfer modification is used to exploit flaws in CNN-based medical imaging systems. Attackers aiming to compromise these systems may purposefully cause mistakes in the diagnosis process, which could result in inaccurate treatment recommendations, misdiagnoses, and eventually patient injury. The basic confidence that patients and healthcare professionals have in automated diagnostic systems is called into question by this possibility of targeted manipulation. Even small system security lapses can have a domino impact on the healthcare ecosystem in an industry where accuracy is critical. Therefore, it is imperative to create strong defenses that guarantee CNN models' continuous dependability, protect them from hostile attacks, and strengthen the general integrity of medical imaging systems.

1.5. Research Objective

1.5.1. Main Objective

The primary objective of this research is to thoroughly investigate the susceptibility of CNN models, particularly those utilized in chest X-ray classification, to STM attacks. These attacks involve subtle alterations to the texture and style of medical images, which can mislead the CNN models into making incorrect classifications, potentially compromising diagnostic accuracy. This research aims to not only assess the extent to which these CNN models are vulnerable to such sophisticated adversarial attacks but also to explore, implement, and evaluate various defense mechanisms. The goal is to identify and develop effective strategies that can enhance the robustness of CNN models against STM attacks, ensuring that these models maintain their reliability and accuracy in clinical settings, even when exposed to adversarial threats.

1.5.2. Sub Objective

- Implementation of STM Attack

Apply the STM technique to chest X-ray datasets using CNN models. This will slightly change the textures and colors of the images in a manner that is almost imperceptible to the human eye but which can mislead the model to make erroneous classifications. This step shall help us evaluate the vulnerability of the model against such adversarial attacks of this particular type.

- Refine CNN Model

Fine-tune CNNs to become better prepared against STM attacks. This would be a fine-tuning for the parameters of the model and training processes to assess and probably improve such adversarial performance, which would go ahead and help boost resilience in models.

- Implementation of Defense Methods

Research and implement different protection techniques against STM-based attacks on CNN models, including data augmentation, which increases the diversity of images shown, and adversarial training, in which the model will be trained with examples of adversaries so it will learn to detect and protect against STM-based attacks.

- Evaluation of Performance Metrics

The performance metrics to be used in evaluating the CNN models under STM attacks will include accuracy, precision, recall, and the F1-score. This would detail how the STM attacks are going to affect the model's diagnostic capability and quantify the extent of performance degradation.

- Validate Model Defense Effectiveness

This involves using CNN models on clean, unaltered datasets versus the performance of models under attack with STM. The result will let one know which of the defense strategies are most effective. This will validate if these defense mechanisms have been successful in maintaining model accuracy and reliability and made it more robust against such adversarial threats.

2. Methodology

2.1. System Diagram

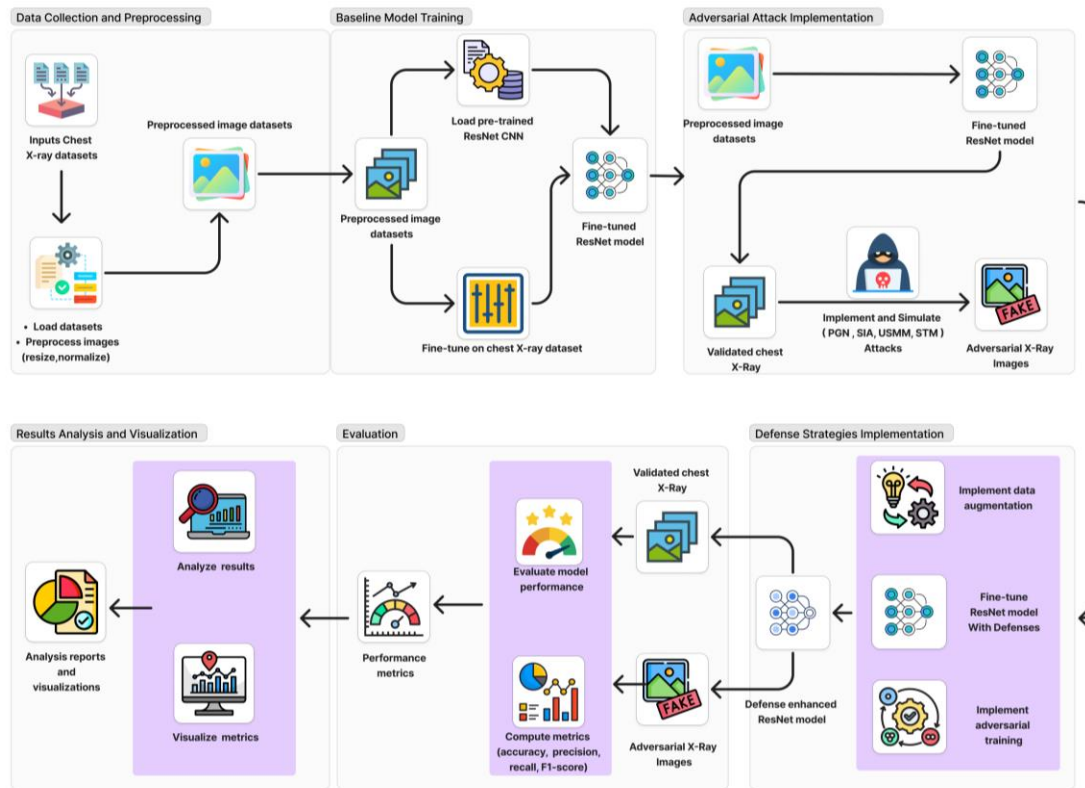


Figure 1: Overall System Diagram

A comprehensive, end-to-end approach for examining and improving the resilience of CNN-based chest X-ray classification models is depicted in the diagram. Raw chest X-ray images are first collected, and they then go through a number of preprocessing stages, including rotation, scaling, normalization, and class balancing. In the end, these preliminary procedures guarantee that the dataset is clear, consistent, and representative, laying a strong basis for the

following modeling stages. This is represented in the diagram as the "Preprocessed image datasets" stage, where the cleaned, unbalanced, and noise-free images are ready for model training.

The preprocessed dataset is then entered into the training module for the baseline model. Here, a CNN model that has already been trained is adjusted to the unique characteristics of medical imaging using the chest X-ray pictures. The network's layers and hyperparameters are adjusted throughout the training phase to attain the best possible performance, and the model is then verified on a subset of the data to create a performance baseline. This phase is essential since it establishes the model's initial ability to categorize the images correctly and prepares the groundwork for subsequent assessments in adversarial settings.

Subsequently, the validated baseline model is subjected to style transfer manipulation (STM) attacks in the adversarial attack implementation phase. In this step, neural style transfer techniques are used to merge the original medical images with a unique design image that features unusual textures and color patterns. In order to guarantee that the adversarial images, despite their perceptual similarity to their original counterparts, include minute perturbations that could deceive the CNN model, the procedure uses an iterative compromise between style loss and content loss. This shift from a validated model to the creation of "attacked images" is depicted in the diagram, which highlights how the adversarial inputs are designed to test the system's weaknesses.

Analyzing and visualizing the results comes after adversarial images are created. Here, metrics including accuracy, precision, recall, and the F1 score are used to assess the CNN model's performance across clean and adversarial datasets. The impact of style transfer changes on diagnostic accuracy can be isolated with the use of visual tools, which produce comprehensive graphs and statistical summaries. In addition to quantifying the performance decline, our approach pinpoints the precise regions of the model that are most vulnerable to adversarial influence.

These findings are then combined in the evaluation step, which thoroughly examines the model's behavior in both adversarial and regular scenarios. The evaluation's feedback is utilized to improve the overall defense mechanisms, fine-tune hyperparameters, and modify attack tactics. The workflow ends with the implementation of defense tactics intended to strengthen the CNN model against these kinds of attacks. Advanced data augmentation, adversarial training which incorporates adversarial samples into the training process—and post-processing methods like JPEG compression and High-Level Representation Guided Denoising are some of these countermeasures. The diagram illustrates how the loop of continuous improvement is closed by reevaluating the improved model following the implementation of these defenses.

Overall, the diagram presents a cohesive methodology that spans data preparation, baseline model training, adversarial attack generation, detailed performance analysis, rigorous evaluation, and the strategic deployment of defense

mechanisms. This sequential yet interconnected process ensures that each phase builds upon the previous one, ultimately leading to a robust investigation into how CNN models for chest X-ray classification can be made resilient against subtle, yet impactful, adversarial threats.

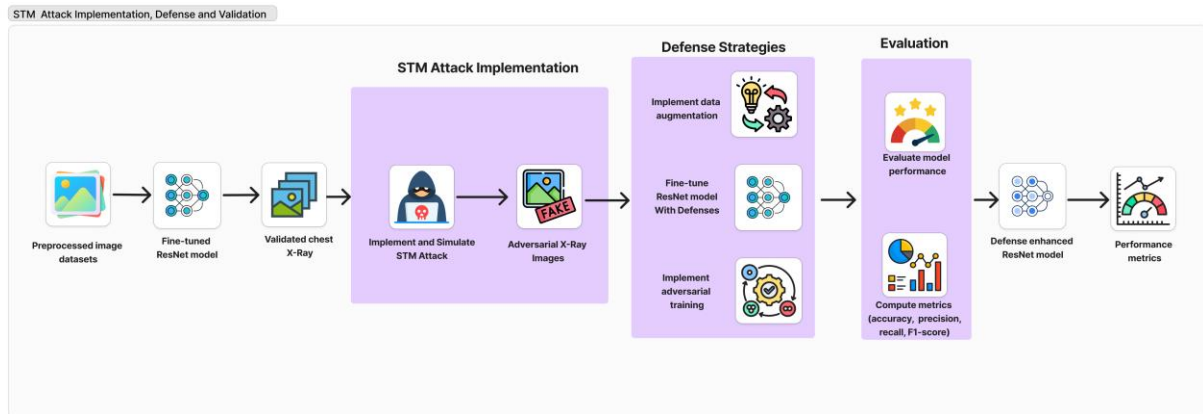


Figure 2: Individual Component System Diagram

The individual component's system diagram shows a simplified procedure for evaluating and improving a chest X-ray classification model based on ResNet50's resistance to style transfer manipulation (STM) attacks. First, the baseline ResNet50 model is refined and validated using preprocessed chest X-ray images. After validation, the model is exposed to STM assaults, which generate adversarial images that can trick the classifier. To increase resistance, a variety of defense techniques are then used, including data augmentation, adversarial training, and model fine-tuning. To close the cycle of continuous improvement, the model is then reassessed using performance metrics (accuracy, precision, recall, and F1-score) to ascertain how effective the defenses are.

2.2. Data Collection

This study is based on a large chest X-ray dataset that initially included 112,000 pictures from 14 different diagnostic categories. This large dataset was carefully selected to represent a variety of pulmonary disorders, offering a wealth of resources for different medical imaging tests.

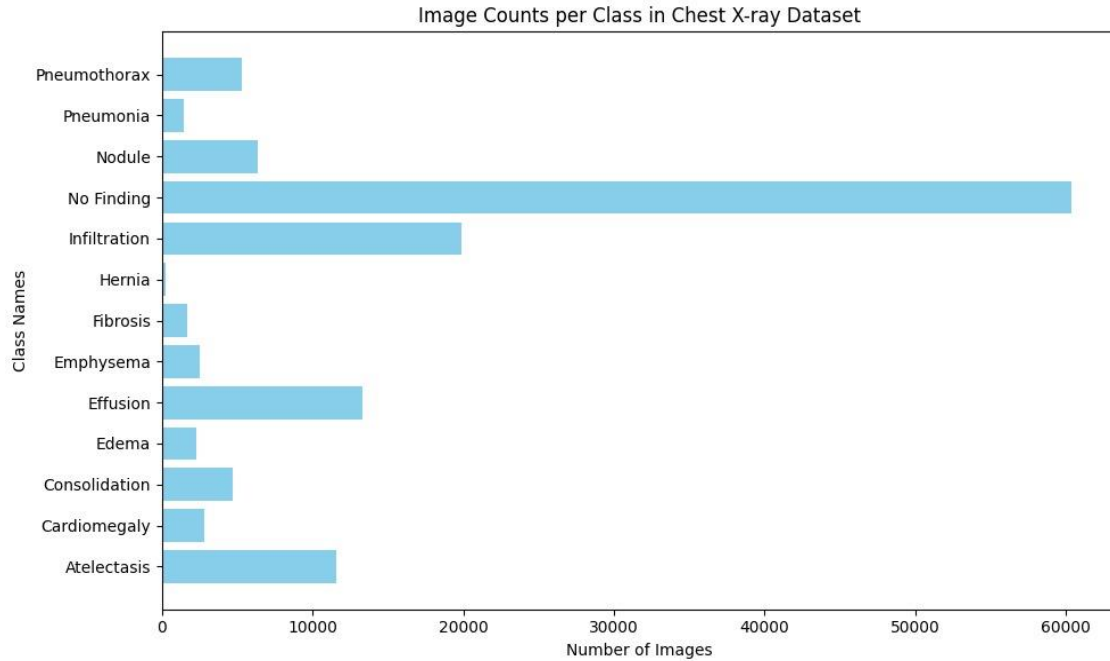


Figure 3: Image Counts per Class of Original Dataset

However, to align more precisely with the objectives of evaluating adversarial robustness in CNN models, we narrowed our focus to three critical classes: normal, pneumonia, and pneumothorax.

Category	Number of Images
Normal	3000
Pneumonia	3000
Pneumothorax	3000

Table 2: Finalized Dataset

For each of these three selected classes, we ensured that our dataset was both representative and balanced by standardizing the number of images per class. Specifically, we aimed for 6,000 images per class, a sample size deemed sufficient to capture the intrinsic variability within each category and robust enough for training deep learning models. In instances where a particular class contained fewer than 6,000 images in the original dataset, we employed a targeted data augmentation technique to normalize the dataset. This augmentation involved rotating the images by 5 degrees a subtle transformation that not only increases the effective sample size but also maintains the inherent diagnostic features present in the images. Such a strategy ensures that each class is equally represented, reducing any potential bias that might arise from class imbalance and further stabilizing the training process

2.3. Selection of the ResNet50 Model

In order to choose the best CNN models for this study, consideration was given to architectures that could both solve frequent training issues and successfully capture the complex, subtle information found in medical images. ResNet50, a well-known residual learning framework, is a prime option that surfaced. By adding shortcut connections that address the vanishing gradient problem a common difficulty in deep neural networks that may delay the learning of complex details where ResNet50's architecture makes it easier to train very deep networks. This feature is especially helpful when examining high-resolution medical images, such chest X-rays, because a correct diagnosis depends on the capacity to accurately identify minor abnormalities. ResNet50 greatly enhances the network's ability to recognize and learn from complex visual patterns by allowing the network to learn residual functions with respect to the layer inputs. This feature is crucial to our study since it immediately improves diagnosis accuracy, which is a critical need in clinical settings.

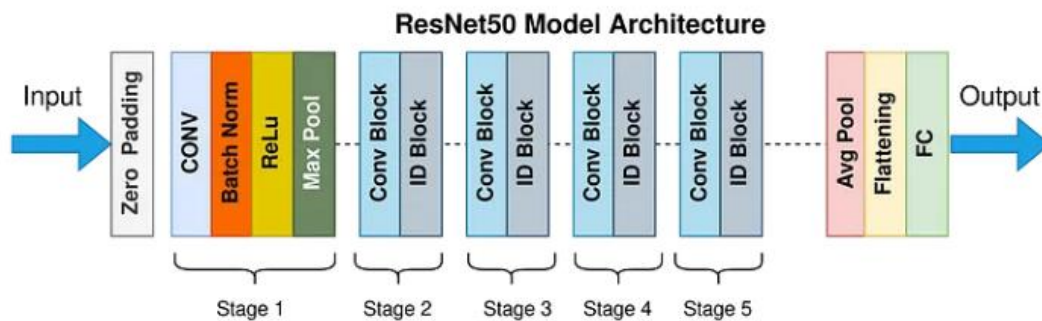


Figure 4: Architecture of ResNet50 Model

ResNet50 was chosen due to its shown performance in a variety of picture identification applications, including those in the medical field, in addition to its sophisticated residual learning capabilities. The CNN architecture must effectively learn from the supplied data without being constrained by deep network training problems since medical images frequently contain overlapping features and subtle texture variations. ResNet50 may concentrate on the key features by avoiding non-essential layers through shortcut connections, which eventually improves model accuracy and resilience to hostile manipulations such style transfer assaults.

In conclusion, ResNet50 was chosen due to its exceptional performance in addressing deep learning problems, especially when it comes to extracting and learning intricate information from medical imaging. This, along with unit consistency and a methodical approach to quantitative analysis, establishes a strong basis for evaluating the model's resilience to complex adversarial threats, such style transfer manipulation.

2.4. Implementation of Style Transfer Attacks

In this study, style transfer attacks are implemented by a complex method that combines stylistic aspects from an external, non-medical design image with the intrinsic content of a medical image. First, a unique design image is carefully chosen because of its unique color patterns and textures, which are uncommon in traditional medical imaging. This choice is crucial since it adds unusual stylistic elements that can gently but successfully alter the goal medical image's appearance. The stylistic elements that will eventually be combined with the original image originate from the selected design image.

The attack uses a neural style transfer technique to merge the two images after the style image has been chosen. In this approach, the medical image's content representation is extracted while the design image's stylistic cues like texture and color are also captured. Style loss and content loss are two conflicting goals that are balanced by optimizing a composite loss function to achieve the fusion. While the content loss component makes sure that the essential diagnostic characteristics and structural integrity of the original medical image are maintained, the style loss component speeds up the integration of new textural and color elements from the design image.

The program uses an iterative gradient descent technique to carry out this balance. The original image's pixel values are gradually changed in each cycle to gradually include the design image's stylistic components. This process keeps going until the adversarial image that is produced—also known as the STM (Style Transfer Manipulation) assaulted image—shows notable stylistic changes while yet being remarkably similar to the original. These slight changes are enough to fool CNN machines into incorrectly classifying the hostile image, which seems almost identical to the human observer.

The final output, as shown in Figure 5, varies sufficiently from the original image to interfere with the learnt features of the model, hence questioning its diagnostic accuracy without raising red flags because of obvious distortions.



Figure 5: STM Attacked image vs. Raw image

The detailed process for executing the style transfer manipulation attack, as described above, is systematically depicted in the algorithm [Appendix 1].

2.5. Implementation of the defense mechanisms

1. Adversarial Training

A mixed dataset of original images and their adversarial altered counterparts is used to train the model in this defense mechanism. The CNN model learns to identify and highlight robust features that hold up even when there is adversarial noise by being exposed to both clean and altered images throughout the training phase. By strengthening the model's resistance, this exposure increases the model's capacity to categorize images accurately and lowers the possibility of misclassification in the presence of subtle hostile alterations.

2. High-Level Representation Guided Denoiser (HGD)

The defense technique incorporates the High-Level Representation Guided Denoiser (HGD) to purify the input images' feature representations. In order to successfully filter out the minute adversarial perturbations brought about by style transfer attacks, HGD aligns high-level semantic features that are taken from the CNN. By removing unnecessary adversarial noise and maintaining important diagnostic information at the feature level, this denoising technique improves the model's overall accuracy and robustness.

3. JPEG Compression

JPEG compression is used as a preprocessing step to reduce hostile artifacts as an extra line of defense. Because JPEG compression is lossy, it naturally minimizes small disturbances, including those caused by style transfer alteration, without significantly compromising the essential diagnostic information of the medical photos. By smoothing out hostile changes, this method serves as a filter that increases the model's resistance to adversarial attacks while preserving the visual integrity required for precise diagnosis.

4. CycleGAN-based defense

By converting stylized (attacked) X-ray pictures from the "adversarial domain" to a "clean domain," this technique seeks to return them to their initial, diagnostic state. A particular kind of Generative Adversarial Network (GAN) architecture called CycleGAN was created to translate images to images without the need for paired training data. The basic notion is that two sets of generators and discriminators are trained together to translate pictures from Domain A (like stylized X-ray images) to Domain B (like clean X-ray images) and vice versa. The technique aims to eliminate style-based distortions while maintaining important medical aspects by imposing cycle uniformity.

5. Feature Consistency Defense

Iteratively modifying an altered image to match reliable references in terms of both content and style is how the defensive mechanism operates. It extracts style representations (by Gram matrices) from a trustworthy style picture and high-level content features from a clean content image using a pre-trained network (like VGG). Gradient descent is then used to minimize a combined loss, which balances variations in style and substance. By adopting the trusted style and preserving the semantic structure of the content image, this approach eventually "corrects" the modified image, negating any adversarial changes.

The process for implementing the Feature Consistency Defense, as described above, is systematically depicted in the following algorithm.

Algorithm 1 Feature Consistency Defense

Input: Content image C , style image S , number of iterations T , style weight λ

Output: Stylized image M

- 1: **Preprocessing:** Resize C and S to a common size
 - 2: **Initialize** $M \leftarrow C$
 - 3: **Extract features using a pretrained VGG**
 - 4: Compute content target: $F_C \leftarrow \text{VGG}(C)$
 - 5: Compute style target: $G_S \leftarrow \text{Gram}(\text{VGG}(S))$
 - 6: **for** $t = 1$ to T **do**
 - 7: Compute features of M : $F_M \leftarrow \text{VGG}(M)$
 - 8: Compute the Gram matrix: $G_M \leftarrow \text{Gram}(F_M)$
 - 9: Compute content loss:

$$\mathcal{L}_{\text{content}} = \|F_M - F_C\|^2$$
 - 10: Compute style loss:

$$\mathcal{L}_{\text{style}} = \|G_M - G_S\|^2$$
 - 11: Compute total loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{content}} + \lambda \mathcal{L}_{\text{style}}$$
 - 12: Update M using gradient descent on $\mathcal{L}_{\text{total}}$
 - 13: **end for**
 - Return:** M
-

2.6. Performance Metrics

In medical imaging, we use a set of thorough performance criteria to assess CNN models' resilience against adversarial type transfer attacks. Both the model's overall accuracy and its capacity to precisely detect positive instances—especially in adversarial settings—are intended to be captured by these metrics. The F1 score, recall, accuracy, and precision are the main metrics employed in this study; each has a matching equation.

The most logical indicator is accuracy, which quantifies the percentage of accurate predictions the model made throughout the whole dataset. The following equation defines it:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where the instances that are correctly recognized as positive are called True Positives (TPs), the instances that are correctly identified as negative are called True Negatives (TNs), the instances that are mistakenly identified as positive are called False Positives (FPs), and the instances that the model misses are called False Negatives (FN). High accuracy in medical imaging means that the model does a good job of identifying images on average, which is important because even a small mistake might have serious clinical consequences.

Precision quantifies the level of the model's accurate predictions. In vital diagnostic systems, it is especially crucial for lowering false alarms. This is how the precision is calculated:

$$Precision = \frac{TP}{TP + FP}$$

According to this formula, accuracy is the proportion of accurately predicted positive cases to all cases that were projected to be positive. A high precision score indicates that the model is likely to be accurate when it predicts a good outcome, such as the existence of pneumonia or pneumothorax, which eliminates the need for needless follow-up operations or treatments.

Sensitivity, also known as recall, measures how well the model can detect every real positive case. It has the following definition:

$$Recall = \frac{TP}{TP + FN}$$

In medical diagnostics, recall is essential since failing to detect a positive case (false negative) might result in conditions going undetected that could have serious repercussions. This metric guarantees that, even in the face of adversarial attempts intended to mask these occurrences, the model is efficient at identifying each instance of a condition from the dataset that is accessible.

The F1 score provides a single metric that balances the accuracy and comprehensiveness of the model's predictions by taking the harmonic mean of precision and recall. It is provided by:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The F1 score provides a more nuanced assessment of the model's performance by combining precision and recall into a single metric. This is especially helpful in situations where the number of positive and negative cases is unbalanced or in adversarial settings where the model's predictive performance may differ between classes.

Together, the metrics offer a robust framework for quantifying the impact of style transfer adversarial attacks on CNN models in medical imaging. They provide critical insights into how adversarial manipulations degrade model performance, ensuring that any proposed defense mechanisms can be accurately assessed in terms of their ability to restore or maintain model accuracy, precision, recall, and overall diagnostic reliability.

2.7. Tools and Technologies

1. Deep Learning Frameworks

- **TensorFlow:** This deep learning framework will be used to implement CNN model and training. It supports the implementation of adversarial attacks, hence necessary research in Style Transfer Manipulation.
- **PyTorch:** The reason for using PyTorch in this research is its dynamic computation graph and usability. It is useful for experimentation and finetuning models. It will be efficient for training and testing the CNN model against STM attacks.

2. Programming Language

- **Python:** This will be the main language used in this research due to its extensive libraries and tools tailored toward machine learning, data manipulation, and visualization. It is also easy and readable, hence appropriate for implementing complex algorithms and handling large datasets.

3. Model Architecture

- **CNN Model:** The CNN model is the chosen architecture for this study. CNN's deep layers make it highly effective in image classification activities, which is crucial for evaluating the impact of STM attacks on medical images.

4. Database Management Tools

- **Pandas:** Pandas for data management, specifically handling chest X-ray datasets in the research. This enables a user to manipulate, clean, and analyze data efficiently. This stage in data preparation goes a long way in training models.
- **NumPy:** NumPy will be used to enhance numerical computations through very powerful tools in array manipulation and mathematical operations. It forms an integral part in handling large datasets and computations required during model training and evaluation.

5. Visualization Tools

- **Matplotlib:** This package will be used for visualization and plotting detailed graphs which support result analysis. It helps in visualizing the performance metrics of the model so that it would be easy to interpret how well it withstands under STM attacks.
- **Seaborn:** It is used in conjunction with Matplotlib for statistical data visualization. This library provides very high-end visualizations that help represent complex relations of data and distributions intrinsic for understanding STM attacks and their defense strategies.

6. Evaluation Metrics

- **Scikit-learn:** This is a ML library that serves to compute most of the evaluation metrics used in this book, such as accuracy, precision, recall, and the F1 score. These are used to benchmark the performance and robustness of the CMM models before and after the application of defense strategies.

7. IDE

- **Google Colab:** Google Colab acts as the IDE for running the experiments. Using Colab provides free usage of its GPUs that are needed for computationally intensive deep learning tasks while training CNN models and simulating STM attacks without requiring a lot of computational resources locally.

8. Hardware and Computer Resources

- **Google Cloud:** For these large-scale computations, the required hardware, computer resources, and other equipment that goes into it, from powerful GPUs to cloud storage, are provided by Google Cloud for this research. This will present the opportunity for ease in handling large datasets and the processing power needed for deep learning.

2.8. Commercialization aspects of the product

The ultimate objective of this project is to turn our academic discoveries into a vendor-ready, scalable, and useful platform that particularly tackles the adversarial flaws in ML models used in medical imaging systems. The four members of our research team each contribute a specialized component centered on a distinct adversarial assault and its accompanying defense technique, allowing us to capitalize on the broad experience of our team in this commercialization plan. Following their refinement, these separate elements will be combined to create a cohesive architecture that serves as the base for a strong adversarial testing and validation platform.

Target Stakeholders:

- **AI Developers and Vendors:** Those building diagnostic models can use our platform to assess vulnerabilities in their systems.
- **Medical Software Companies:** Firms integrating CNN-based radiology tools will benefit from enhanced security assessments.
- **Healthcare Institutions:** Hospitals and clinics deploying deep learning solutions can utilize the platform to ensure their models are robust and reliable prior to clinical implementation.

User Benefits:

- Provides a controlled, modular, and repeatable testing environment.
- Empowers healthcare-focused AI developers to proactively harden their models before deployment.
- Enhances the security, trustworthiness, and regulatory compliance of medical imaging systems.

Our commercialization strategy also details a number of upcoming improvements intended to expand the platform's functionality:

- **Real-Time API Integration:** Facilitating smooth, real-time communication with current AI pipelines for ongoing robustness assessment and model monitoring.
- **Modality Expansion:** Adding support for further imaging modalities, such as CT and MRI, to increase the platform's versatility in relation to different clinical diagnostic instruments.
- **Advanced Visualization and Security Scoring:** Including thorough dashboards that offer security scoring and real-time performance indicators, which promote compliance and help clinical settings make well-informed decisions.

The overall goal of this commercialization strategy is to provide the first adversarial validation framework designed especially for the medical imaging industry. Our technology is expected to significantly improve healthcare by bridging the gap between cutting-edge AI research and workable cybersecurity solutions. This will ensure that diagnostic models are secure, reliable, and completely compliant with regulations prior to being used in crucial clinical procedures.

2.9. Testing & Implementation

1. Implementation of ResNet50 Model

```
# Split the data into 80% training and 20% testing sets
train, test = train_test_split(data, test_size=0.2, random_state=42)

# Split the training data into 75% training and 25% validation sets.
train, valid = train_test_split(train, test_size=0.25, random_state=42)
```

Figure 6: Code Snippet of Dataset Split

This code snippet divides the dataset into three parts:

- **Training Set (60%):** First, splits the data into 80% training and 20% testing. Then, it further splits the training portion (80%) into 75% training and 25% validation, resulting in 60% of the original data being used for final training.
- **Validation Set (20%):** From the 80% training split, 25% is reserved for validation.
- **Testing Set (20%):** The initial split directly allocates 20% of the dataset for testing.

```
# Data Augmentation
train_gen = train_datagen.flow_from_dataframe(
    dataframe=train,
    x_col='Filepath',
    y_col='Label',
    target_size=(224, 224),
    class_mode='categorical',
    batch_size=32,
    shuffle=True,
    seed=42 # Ensure reproducibility
)

valid_gen = train_datagen.flow_from_dataframe(
    dataframe=valid,
    x_col='Filepath',
    y_col='Label',
    target_size=(224, 224),
    class_mode='categorical',
    batch_size=32,
    shuffle=False,
    seed=42
)

test_gen = test_datagen.flow_from_dataframe(
    dataframe=test,
    x_col='Filepath',
    y_col='Label',
    target_size=(224, 224),
    class_mode='categorical',
    batch_size=32,
    shuffle=False # No shuffling for testing
)
```

Figure 7: Code Snippet of Data Augmentation

The above code snippet uses **Keras**'s `flow_from_dataframe()` method to generate batches of images from dataframes for training, validation, and testing. Each generator is configured with key parameters:

Key Parameters:

- `x_col='Filepath'` and `y_col='label'` specify image file paths and labels.
- `target_size=(224, 224)` resizes all images to a consistent size compatible with ResNet (224×224).
- `class_mode='categorical'` indicates multi-class classification.
- `batch_size=32` processes data in mini-batches of 32.
- `shuffle=True` randomizes the order of images each epoch.
- `seed=42` ensures reproducibility of the random shuffling.

Overall, these generators streamline how images are loaded, processed, and augmented (if specified) for each phase (train, validation, test), supporting an organized and reproducible training workflow.

```
# Build custom classification head
x = layers.GlobalAveragePooling2D()(base_model.output)      # pool to vector
x = layers.Dense(128, activation='relu')(x)                  # 1st dense layer
x = layers.Dropout(0.5)(x)                                   # dropout for regularization
outputs = layers.Dense(3, activation='softmax')(x)           # final output layer for 3 classes
model = models.Model(inputs=base_model.input, outputs=outputs)
```

Figure 8: Code Snippet of Layers Added

This snippet adds a custom classification head to an existing convolutional base model. First, it applies `GlobalAveragePooling2D()` to transform the output feature maps of the base model into a 1D vector. Next, a `Dense(128, activation='relu')` layer introduces learnable parameters for further feature processing, followed by `Dropout(0.5)` to reduce overfitting. Finally, a `Dense(3, activation='softmax')` layer produces class probabilities for 3 distinct categories. The Keras Functional API then wraps these layers with the base model's input and new output to form the final model.

```
# Compile the model (phase 2) with a lower learning rate for fine-tuning
model.compile(optimizer=optimizers.Adam(learning_rate=1e-4),
              loss='categorical_crossentropy',
              metrics=['accuracy'])

# Early stopping: stop if validation loss doesn't improve for 3 epochs (restore best model)
early_stop = EarlyStopping(monitor='val_loss', patience=3, restore_best_weights=True, verbose=1)

# Reduce LR on Plateau: reduce learning rate if validation loss plateaus
reduce_lr = ReduceLROnPlateau(monitor='val_loss', factor=0.1, patience=2, verbose=1)
```

Figure 9: Code Snippet of Callbacks and Early stopping

In this phase of training, the model is recompiled with a lower learning rate ($1e-4$) for fine-tuning. Two callbacks are set up for EarlyStopping and ReduceLROnPlateau.

2. Implementation of Style Transfer Manipulation Attack

The below image depicts the result of performing the STM by blending the style image at 100% opacity into the original chest X-ray.

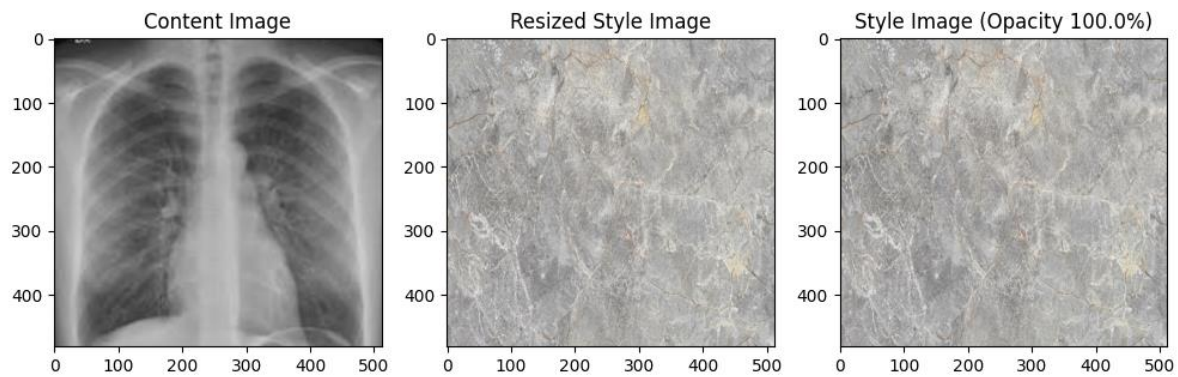


Figure 10: STM Attack at 100% Opacity

The below image depicts the result of performing the STM by blending the style image at 50% opacity into the original chest X-ray.

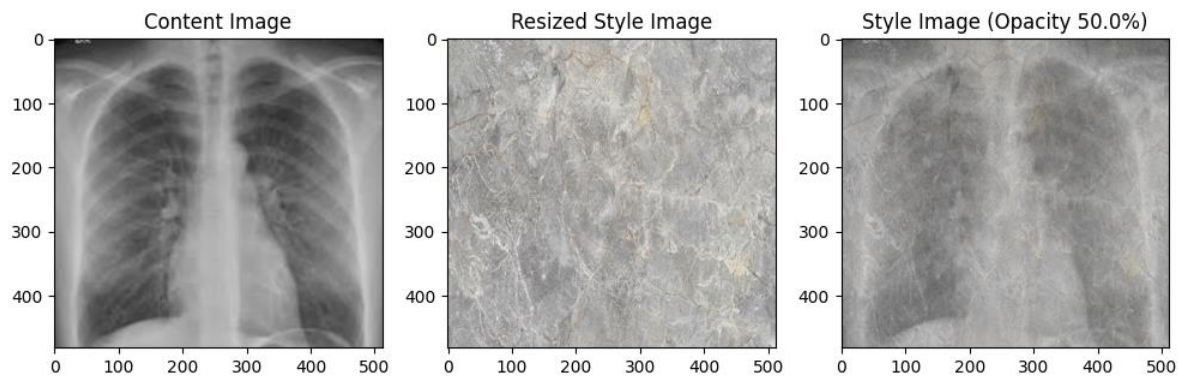


Figure 11: STM Attack at 50% Opacity

The below image depicts the result of performing the STM by blending the style image at 30% opacity into the original chest X-ray.

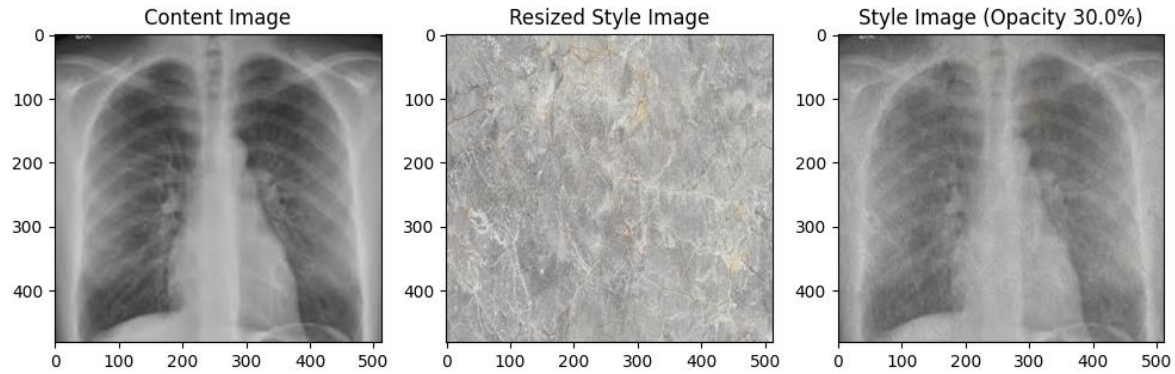


Figure 12: STM Attack at 30% Opacity

The below image depicts the result of performing the STM by blending the style image at 10% opacity into the original chest X-ray.

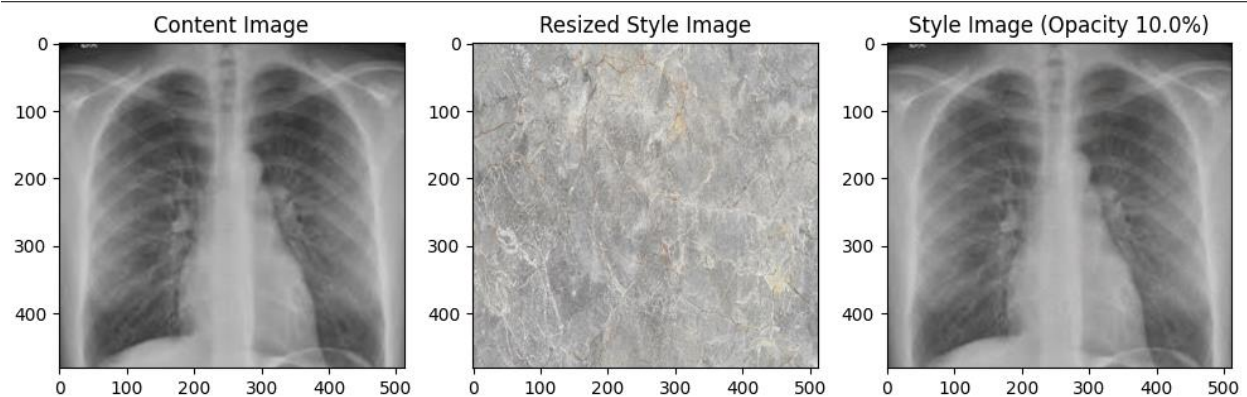


Figure 13: STM Attack at 10% Opacity

```
# Reduce opacity of style image
alpha = 1.0 # Adjust opacity level (0.0 = fully transparent, 1.0 = fully visible)
blended_style_image = (alpha * style_image) + ((1 - alpha) * content_image)
```

Figure 14: Code Snippet of 10% Opacity

Opacity Reduction: To ensure that the adversarial modifications are not easily noticeable, you reduce the opacity (or intensity) of the style image. This is usually achieved by assigning a small blending factor (often called alpha) to the style image. The lower the alpha value, the less visible the changes will be when the style image is superimposed on the original image.

3. Detecting the Style Transfer manipulation Attack

The original and attacked/styled images are scaled in the detecting procedure to guarantee that their dimensions match. After that, structural characteristics are extracted using the Canny edge detector, producing edge maps for both

pictures. An "edge score" that measures structural changes is generated by averaging the absolute difference of these edge maps across all pixels. The Structural Similarity Index (SSIM), which measures perceptual degradation, is calculated simultaneously between the grayscale versions of the images. A score around 1 indicates great similarity, whereas a value below 1 indicates considerable changes. The edge score and SSIM score are compared to predetermined thresholds in order to determine the detection decision; if the edge score is higher than the threshold or the SSIM score is lower than the threshold, an attack is identified. Furthermore, frequency domain analysis based on FFT are carried out and displayed to further expose manipulation-induced variations. This dual method improves the accuracy of identifying modest image attacks by capturing both localized structural alterations and overall image degradation.

The key formulas used in the detection mechanism are explained in Appendix 2.

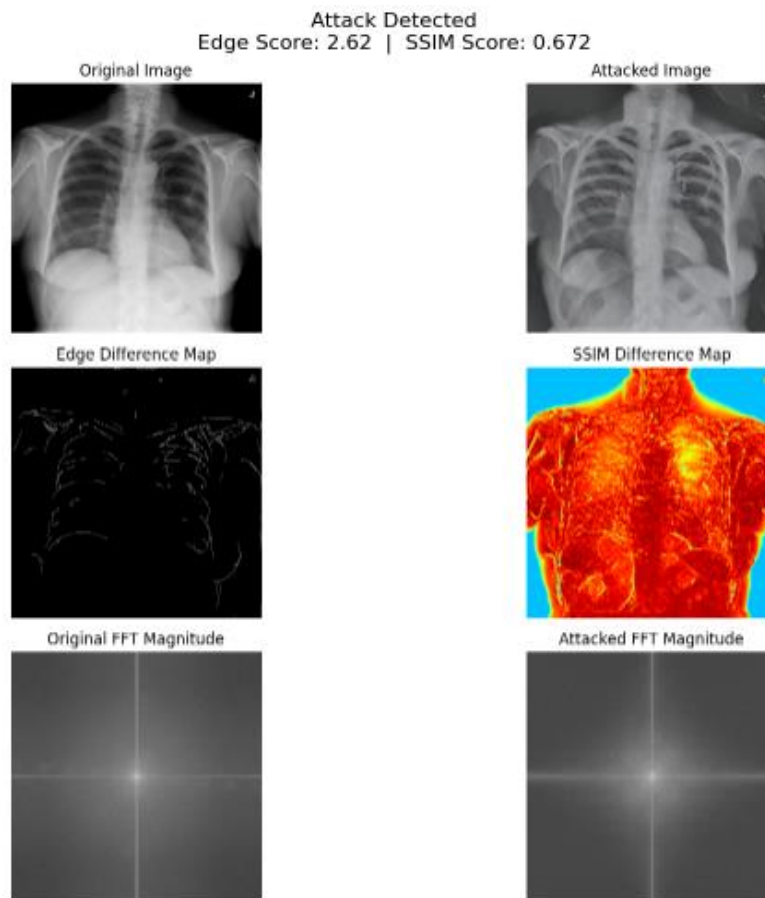


Figure 15: Detection of Normal Category

The detection system flagged the image because the overall similarity between the original and suspected image was much lower than expected (SSIM of 0.672), even though the changes in edge details were minimal. This suggests that differences in brightness, contrast, or global structure are significant enough to trigger an "Attack Detected" result.

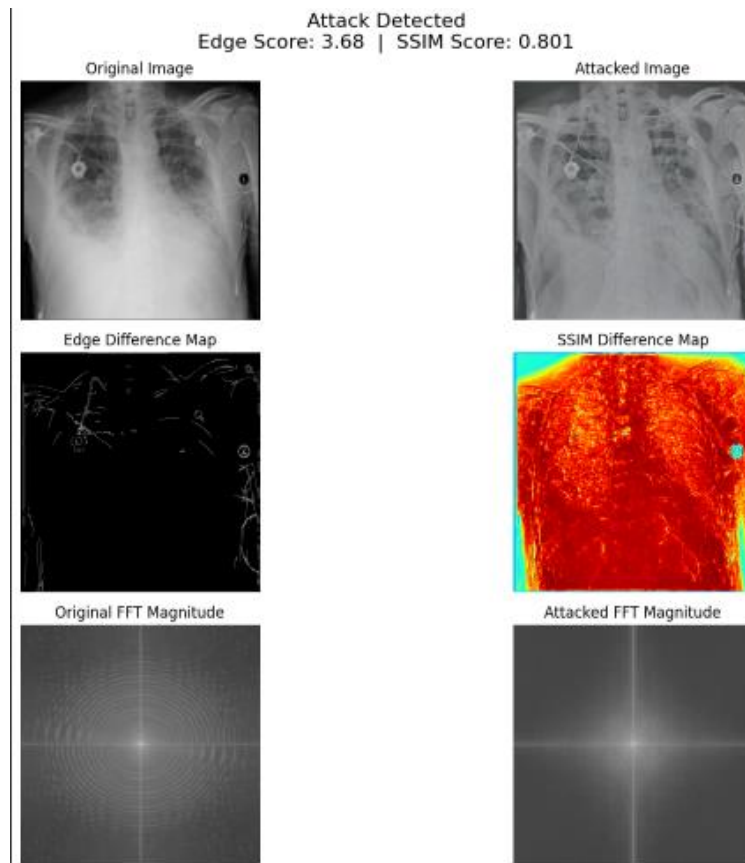


Figure 16: Detection of Pneumonia Category

The system labeled this pair “Attack Detected” mainly because the SSIM (0.801) is below the threshold of 0.95, indicating noticeable structural or brightness changes across the image. Although the edge difference map score (3.68) isn’t extremely high, it still reflects some alterations, and the bright regions in the SSIM difference map confirm that differences are fairly widespread, prompting the detector to flag the image.

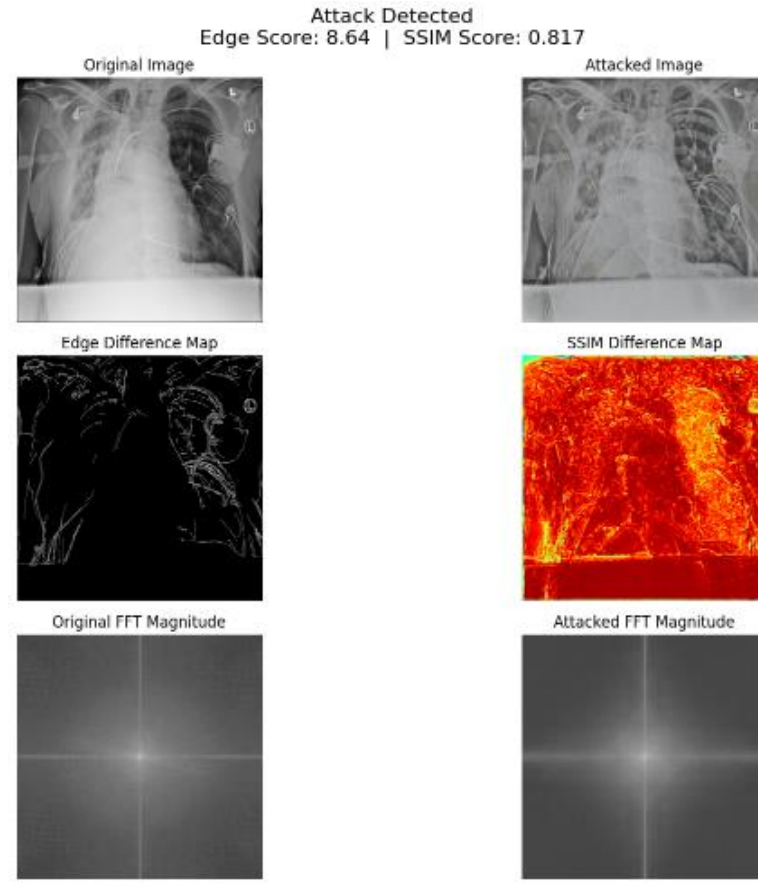


Figure 17: Detection of Pneumothorax Category

Here, both the high edge difference score (8.64) and the relatively low SSIM score (0.817) point to significant changes in structure and brightness. The bright areas in the SSIM map confirm these widespread alterations, causing the system to label the image as "Attack Detected."

4. Implementation of Defense Mechanism

1. Adversarial Training:

This method involved training the CNN model using both the original clean images and the adversarially manipulated images. By exposing the model to a diverse set of examples during training, it learned to distinguish between genuine diagnostic features and adversarial perturbations, effectively improving its robustness.

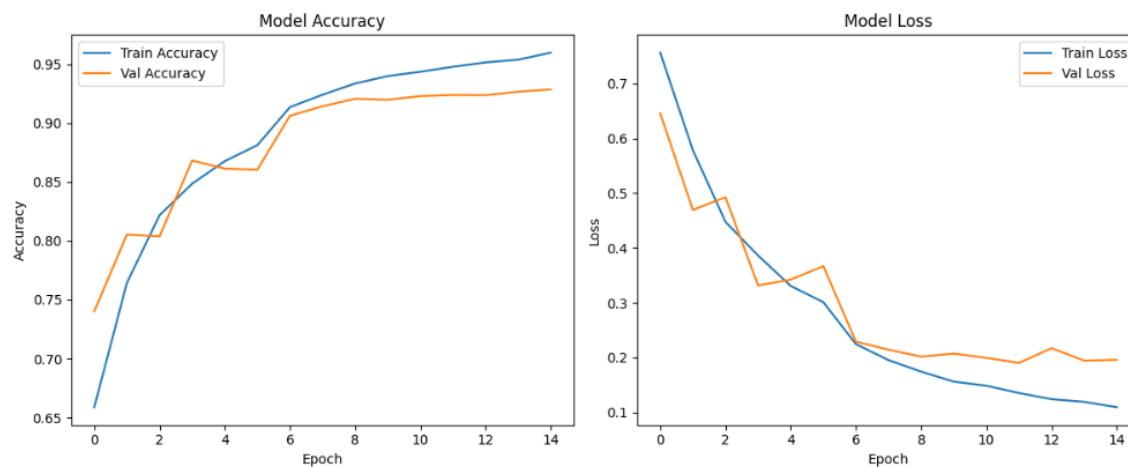


Figure 18: Training and Validation accuracy of Adversarial Training

	precision	recall	f1-score	support
Normal	0.88	0.91	0.90	2441
Pneumonia	0.97	0.99	0.98	2359
Pneumothorax	0.92	0.88	0.90	2400
accuracy			0.92	7200
macro avg	0.92	0.92	0.92	7200
weighted avg	0.92	0.92	0.92	7200

Figure 19: Classification Metrics of Adversarial Training

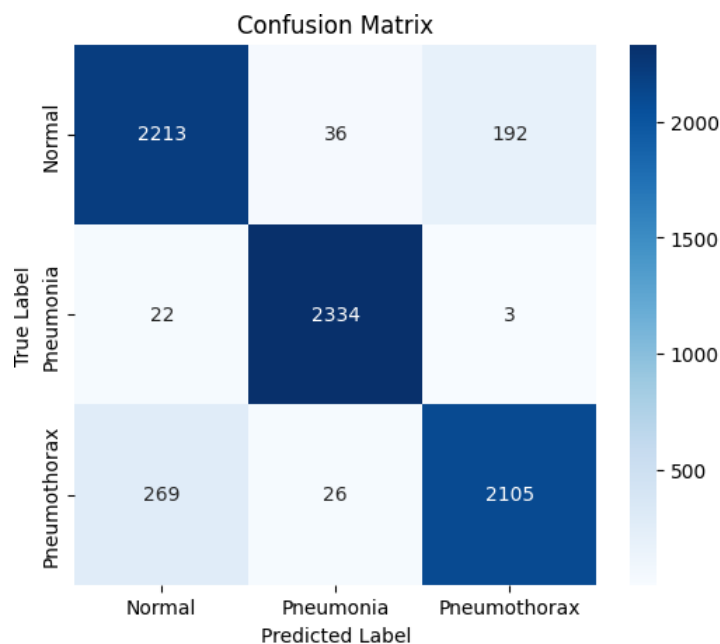


Figure 20: Confusion Metrics of Adversarial Training

2. High-Level Representation Guided Denoiser (HGD):

Although HGD was expected to remove adversarial noise by aligning high-level semantic features, applying it to chest X-ray images led to an unintended consequence. The denoising process excessively smoothed the images, resulting in nearly grey outputs that lacked the critical diagnostic details necessary for accurate analysis.

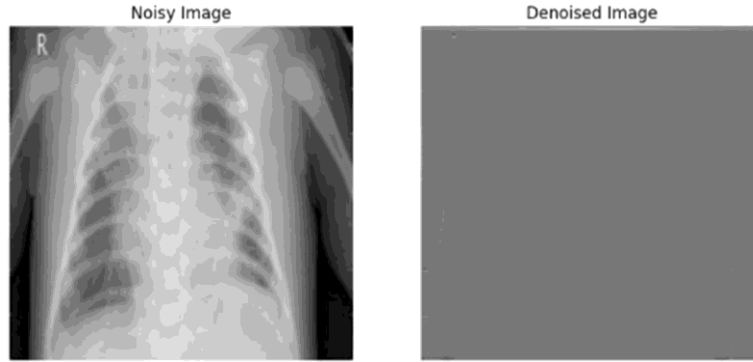


Figure 21: Output of HGD Denoising Technique

3. JPEG Compression:

Similarly, JPEG compression was tested as a technique to filter out high-frequency adversarial artifacts. However, due to the inherently subtle details in X-ray images, this approach also degraded the images further, compromising their diagnostic quality rather than enhancing robustness.

4. CycleGAN-based defense

To implement the CycleGAN model, a dataset of original chest X-ray images (representing the clean domain) and their corresponding style-transferred versions (representing the stylized/adversarial domain) would be prepared. The CycleGAN training process involves:

- **Generator A→B and Generator B→A:**

Two generators are employed: one learns to translate the adversarially stylized X-ray images into clean X-rays, while the other converts clean X-rays back into stylized versions.

- **Discriminator A and Discriminator B:**

Each generator has a dedicated discriminator that evaluates how realistic the generated images are in their respective domains. The discriminators help guide the generators to produce outputs that closely match the characteristics of the target domain.

- **Cycle Consistency Loss:**

A key component of CycleGAN is the cycle consistency loss, which requires that an image translated from Domain A to B and then back again to A should closely resemble the original image. This constraint is meant to ensure that the medical content remains intact during the transformation process.

- **Adversarial Loss:**

Standard GAN adversarial losses for both domains drive the generators to produce outputs that fool the discriminators, pushing the model to produce realistic “clean” X-ray images.

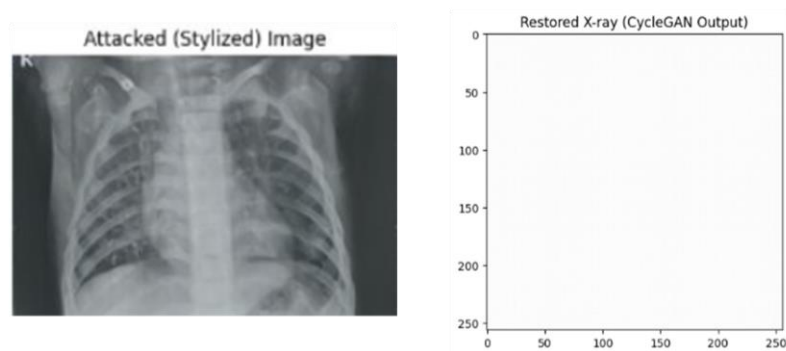


Figure 22: Output of CycleGAN Defense Technique

5. Feature Consistency Defense

The following steps show the implementation of the feature consistency defense technique

1. 1. Preprocessing

2. Resize C and S to the same size.

- Ensures both images match the input dimensional requirements of the network.
- Common practice is to fix the images at a manageable resolution (e.g., 256×256 or 512×512).
- Initialize $M \leftarrow C$
- The image M (which will be iteratively updated) is initially set to the content image.

Starting from C makes sense when you aim to preserve the semantic content. Alternatively, one could start from another initialization (e.g., random noise), but the algorithm here begins from the content image.

3. Extract Features Using a Pretrained VGG

1. Content Target $F_C \leftarrow \text{VGG}(C)$

- Pass the content image C through a truncated VGG network
- Capture the **content feature** representation, denoted F_C .

- Typically, these are features from a higher convolutional layer, where the network focuses on larger-scale structures (e.g., shapes, objects).
-
- 2. **Style Target $G_S \leftarrow \text{Gram}(\text{VGG}(S))$**
 - Pass the style image S through the same VGG network (potentially capturing one or multiple layers, depending on implementation).
 - Compute the **Gram matrix** of those feature maps, denoted G_S .
 - The Gram matrix measures how feature channels correlate with each other, effectively capturing the “style” (textures, colors, patterns).

4. Iterative Optimization

We run a loop from $t=1$ to T . Each iteration nudges M to reduce the total loss, which is composed of both content and style components.

For $t=1$ to T^{**} :

1. **Compute features of M :**

$FM \leftarrow \text{VGG}(M)$

- This extracts the current feature representation of the image M .

2. **Compute Gram matrix GM :**

$GM \leftarrow \text{Gram}(FM)$

- Represents the style information of the current image M by correlating feature channels.

3. **Compute content loss:**

$\lambda_{\text{content}} = \|FM - FC\|^2$

- Measures how different the content features of M are from those of C .
- A smaller value means the image M is closer in content structure to C .

4. **Compute style loss:**

$\lambda_{\text{style}} = \|GM - GS\|^2$

- Measures the difference between M 's style representation and the style target from S .

- A smaller value means $\{M\}$ has similar textures, colors, and patterns to S .

5. **Compute total loss:**

$$\lambda_{\text{total}} = \lambda_{\text{content}} + \lambda_{\text{style}}.$$

- Balances content fidelity (λ_{content}) with style similarity (λ_{style}).
- λ controls how strongly style influences the final image relative to content.

6. **Update M using gradient descent on λ_{total} :**

- Calculate the gradients of λ_{total} w.r.t. the pixels of M .
- Perform a gradient descent (or an Adam optimizer update) step:

$$M \leftarrow M - \eta \nabla_M \lambda_{\text{total}},$$

where η is the learning rate.

- This step iteratively corrects M to reduce λ_{content} and λ_{style} , moving M closer to having C 's content S 's style.

End For

5. Return M

- After T iterations, the algorithm outputs the updated image M .
- This final image M ideally preserves the main semantic structure of the content image C while exhibiting the style of S .

3. Results and Discussion

3.1. Results of the Baseline ResNet50 Model

The below figure depicts the training and validation accuracy of the baseline ResNet50 model (before unfreezing deeper layers)

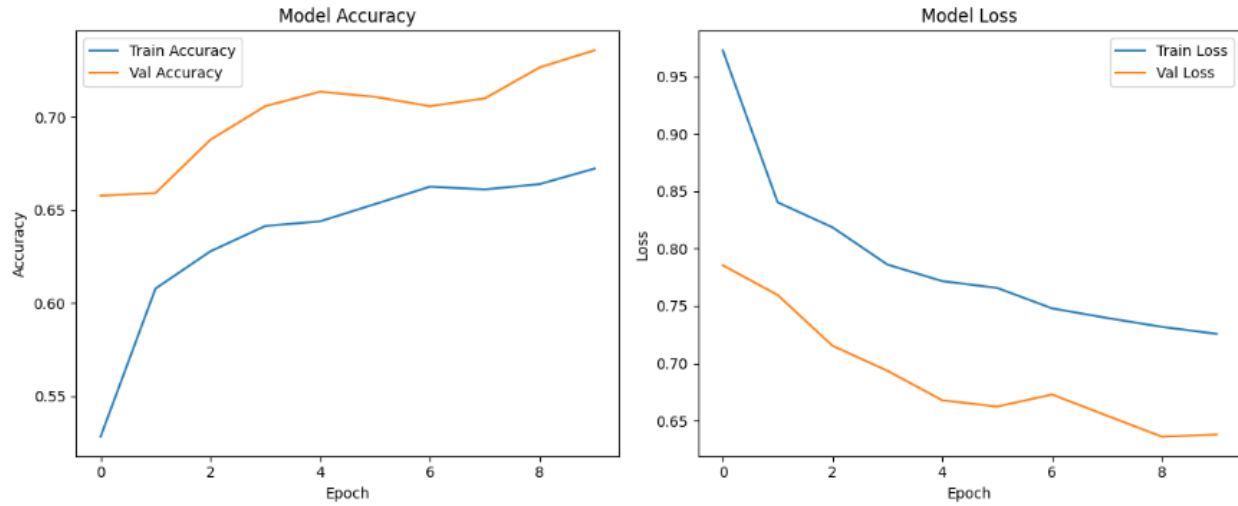


Figure 23: Baseline Model Training (Before unfreezing layers)

The below figure depicts the training and validation accuracy of the baseline ResNet50 model (after unfreezing deeper layers)

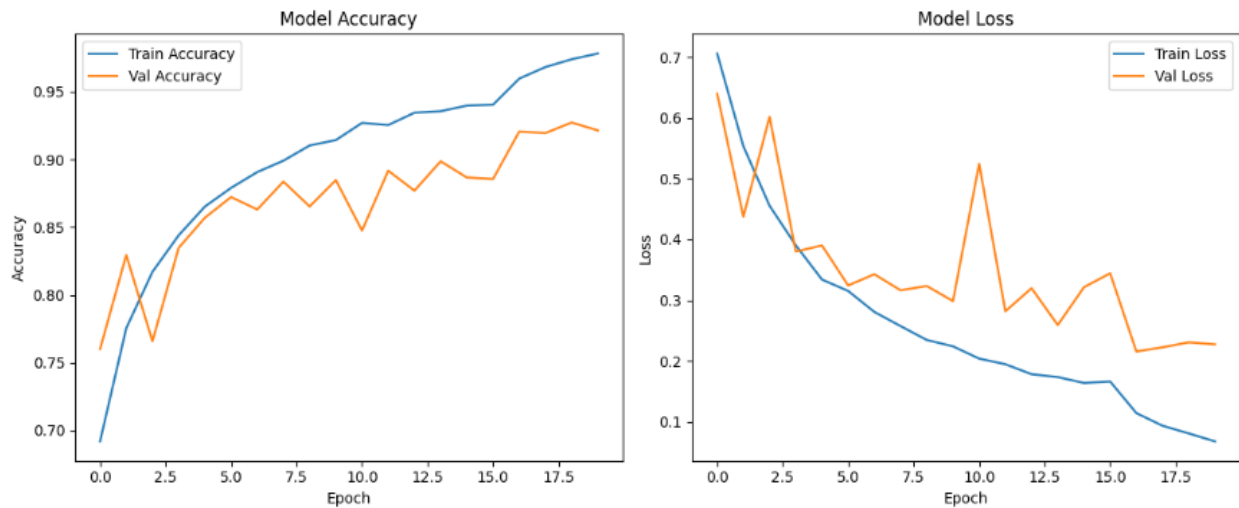


Figure 24: Baseline Model Training (After unfreezing layers)

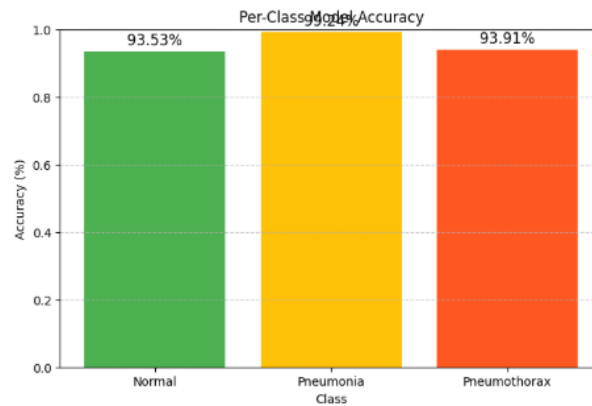


Figure 25: Per-Class Accuracy of Baseline Model

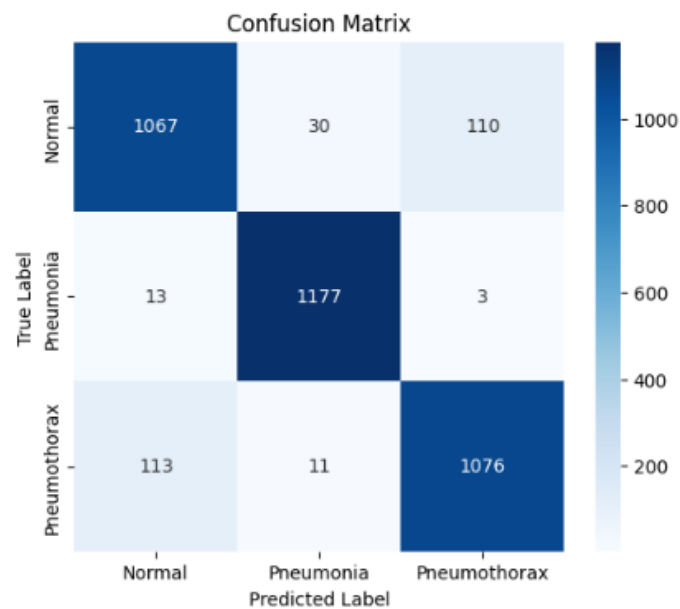


Figure 26: Confusion Metrics of baseline Model

The model's ability to differentiate between the three classes Normal, Pneumonia, and Pneumothorax based on anticipated versus real labels is displayed in this confusion matrix. The model performs well overall, as evidenced by the diagonal cells (1067 Normal, 1177 Pneumonia, and 1076 Pneumothorax), which reflect accurate classifications. Nonetheless, there are significant misclassifications, particularly between Normal and Pneumothorax (113 cases), which implies that these two groups occasionally have characteristics in common. On the other hand, the most correctly identified class is pneumonia, with just a little amount of uncertainty (13 Normal, 3 Pneumothorax). Overall, even while the model correctly classifies most photos, it could be improved for all classes and the Normal-Pneumothorax misunderstanding might be lessened with more development.

	precision	recall	f1-score	support
Normal	0.89	0.88	0.89	1207
Pneumonia	0.97	0.99	0.98	1193
Pneumothorax	0.90	0.90	0.90	1200
accuracy			0.92	3600
macro avg	0.92	0.92	0.92	3600
weighted avg	0.92	0.92	0.92	3600

Figure 27: Classification Metrics of Baseline Model

These classification metrics indicate strong overall performance, with an accuracy of 0.92 across the three classes. Pneumonia is most accurately identified, evidenced by its near-perfect precision (0.97), recall (0.99), and F1-score (0.98). The Normal and Pneumothorax classes also exhibit solid metrics, each maintaining an F1-score near or above 0.89. The macro and weighted averages of 0.92 underscore the model's balanced performance across the class distribution, though some room remains for further improvements, particularly in distinguishing Normal from Pneumothorax cases.

3.2. Research Findings

3.2.1. Evaluation of clean images (Before the Attack)

With a 95.51% overall accuracy on the clean chest X-ray dataset, the model performs exceptionally well in identifying pneumonia (precision: 0.98, recall: 0.99, F1-score: 0.99). Despite achieving high F1-scores of 0.94 for both the Normal (precision: 0.92, recall: 0.95) and Pneumothorax (precision: 0.96, recall: 0.92) classes, the confusion matrix shows that the most frequent misclassifications are between Normal and Pneumothorax (244 Normal cases labeled as Pneumothorax, and 431 Pneumothorax cases labeled as Normal). The model is sometimes misled by these minor overlapping features. The nearly equal precision, recall, and F1-scores for each of the three classes demonstrate the model's strong diagnostic capacity before any hostile interference, albeit these flaws.

```

==== Clean Images Evaluation ====
Accuracy: 95.51%
Classification Report:

```

	precision	recall	f1-score	support
Normal	0.92	0.95	0.94	6000
Pneumonia	0.99	0.99	0.99	6001
Pneumothorax	0.96	0.92	0.94	6000
accuracy			0.96	18001
macro avg	0.96	0.96	0.96	18001
weighted avg	0.96	0.96	0.96	18001

Figure 28: Classification Metrics before the Attack

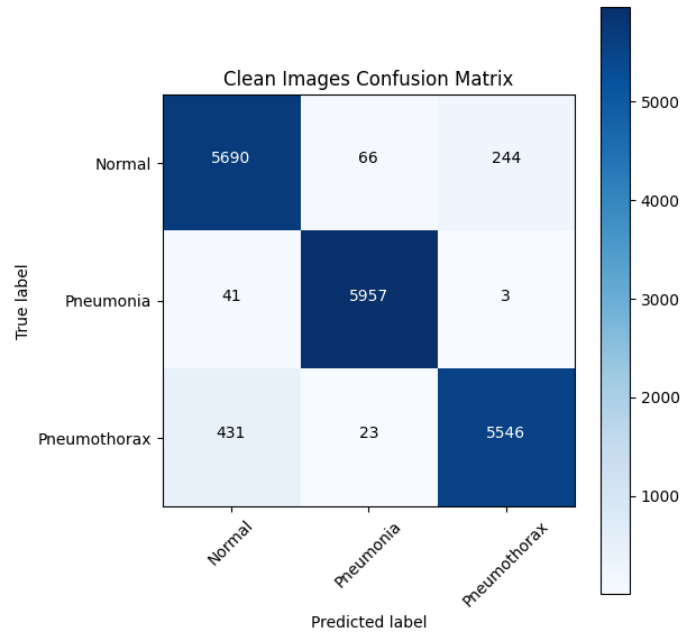


Figure 29: Confusion Matrix before the Attack

3.2.2. Evaluation of Styled Images (After the Attack)

The total accuracy of the model falls drastically from over 95% to 38.63% after the style transfer manipulation is applied. The classification report shows a sharp decline in performance, with Pneumonia's recall falling to 0.02—a sign that the model rarely properly recognizes pneumonia images (91 out of 6000). As a result of the style manipulation's capacity to skew important diagnostic characteristics, many cases of pneumonia and normal illness are mistakenly categorized as pneumonia. Despite the fact that pneumothorax is still comparatively easily identified (recall 0.82), the significant misunderstanding between classes highlights how successfully the STM attack compromises the diagnostic accuracy of the model.

```

Accuracy: 38.63%
Classification Report:

```

	precision	recall	f1-score	support
Normal	0.37	0.32	0.34	6000
Pneumonia	0.99	0.02	0.03	6000
Pneumothorax	0.39	0.82	0.53	6000
accuracy			0.39	18000
macro avg	0.58	0.39	0.30	18000
weighted avg	0.58	0.39	0.30	18000

Figure 30: Classification Metrics after the Attack

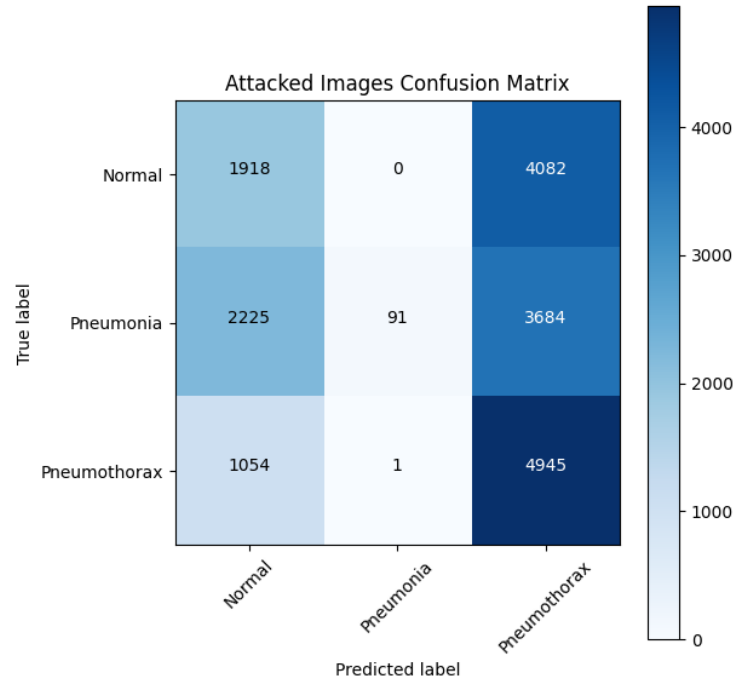


Figure 31: Confusion Matrix after the Attack

3.2.3. Evaluation of Defenses Images (After implementing Defense)

The total accuracy increases to 50% from the much lower accuracy shown under the raw STM attack after feature consistency and adversarial training are applied as defenses. The classification report and confusion matrix, however, reveal significant differences in the treatment of each class. The model has reasonable success in detecting Normal and Pneumothorax, with recalls of 0.80 and 0.65, respectively. However, it has significant difficulty in identifying Pneumonia, accurately identifying only 1 out of 20 samples (recall of 0.05). This discrepancy shows that although the defense techniques lessen some of the attack's stylistic distortions, they are unable to sufficiently restore the essential characteristics required to identify pneumonia, underscoring the necessity for additional defense strategy improvement.

```

===== Defended Images Evaluation =====
Accuracy: 50.00%
Classification Report:

```

	precision	recall	f1-score	support
Normal	0.38	0.80	0.52	20
Pneumonia	1.00	0.05	0.10	20
Pneumothorax	0.76	0.65	0.70	20
accuracy			0.50	60
macro avg	0.72	0.50	0.44	60
weighted avg	0.72	0.50	0.44	60

Figure 32: Classification Metrics after Defense

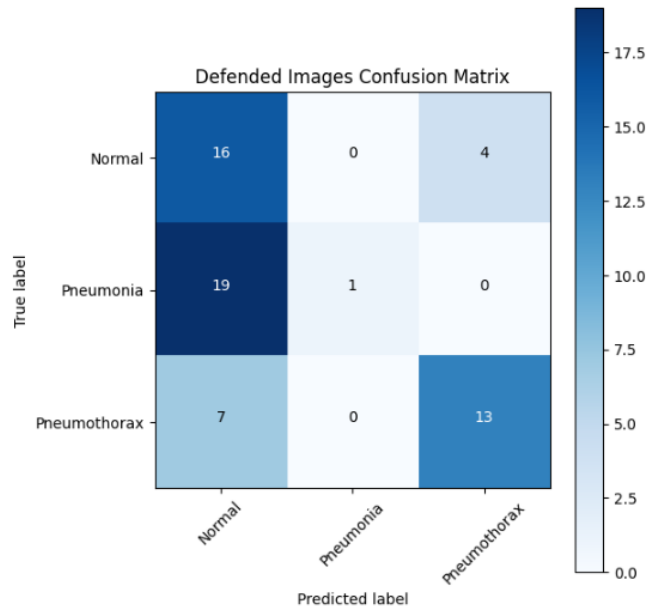


Figure 33: Confusion Matrix after Defense

3.3. Discussion

Instance	Before Attack	After Attack	After Defense
Accuracy	96%	39%	50%

Table 3: Summary of Findings

According to the summary, on clean, unaltered chest X-ray images, the baseline CNN model had a high accuracy of 96%. The accuracy sharply decreased to 39% following the style transfer manipulation attack, demonstrating the negative impact of adversarial perturbations on model performance. The accuracy partially recovered to 50% after defense measures, such as feature consistency and adversarial training, were applied. This suggests that although the defenses provide some mitigation, they do not completely restore the model's initial performance.

4. Summary of Each Student's contribution

Name	Registration Number	Functions
Nihila Premakanthan	IT21197550	<ul style="list-style-type: none">• Data Acquisition and Preprocessing• Baseline Model (ResNet50) Training and Implementation• Implementation of Style Transfer Manipulation Attack• Implementation of Detection Mechanism for STM Attack• Implementation of Defense Mechanisms• Comparative Analysis(Before attack, After attack, After Defense)• Assessing the Robustness of ResNet50 Model against STM Attack

Table 4: Individual Contribution

5. Conclusion

In conclusion, this research has demonstrated that although CNN-based diagnostic systems, such as those built on the ResNet50 architecture, can achieve remarkable accuracy on clean chest X-ray images, they remain highly susceptible to adversarial style transfer manipulation attacks. Our experiments revealed that, despite an initial accuracy of 96% on unaltered images, the application of style transfer manipulation attacks caused a dramatic drop in performance to 39%. This significant degradation underscores the critical vulnerability of these models when adversarial perturbations are introduced.

The study explored a range of defense mechanisms designed to counteract these adversarial effects, including adversarial training, High-Level Representation Guided Denoiser (HGD), JPEG compression, CycleGAN-based restoration, and a feature consistency approach. Among these, adversarial training provided the most feasible improvements, partially recovering the model's performance to an accuracy of 50%. However, the other defense strategies, while conceptually promising, either overly smoothed the diagnostic details or introduced unwanted artifacts, thereby failing to adequately preserve the critical information required for clinical interpretation.

These findings not only highlight the formidable challenge of protecting medical imaging systems from adversarial threats but also illuminate a significant research gap in this domain. The limited effectiveness of the current defense techniques calls for further exploration into more robust, medically tailored defense strategies that can maintain diagnostic integrity even under adversarial conditions. Future research should focus on the development of innovative methodologies that can effectively neutralize adversarial perturbations without compromising the sensitive diagnostic features inherent in medical images, as well as on integrating these defenses into real-time clinical workflows.

6. References

- [1] F. S. H. L. Y. L. L. W. W. F. X. W. Zhijin Ge, Improving the Transferability of Adversarial Examples with Arbitrary Style Transfer, 2023.
- [2] X. Y. a. K. H. Zhilu Zhang, Attacking Sequential Learning Models with Style Transfer Based Adversarial Examples, IOP Publishing, 2021.
- [3] A. Thangaraju and C. Merkel, Exploring Adversarial Attacks and Defenses in Deep Learning, Bangalore, India: 2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), 2022.
- [4] J. a. D. X. a. H. X. a. Y. F. a. T. Q. a. C. T.-S. Tang, Adversarial Training Towards Robust Multimedia Recommender System, IEEE Transactions on Knowledge and Data Engineering, 2020.
- [5] M. L. ., Y. D. T. P. X. H. ., J. Z. Fangzhou Liao*, Defense Against Adversarial Attacks Using High-Level Representation Guided Denoiser, Salt Lake City, UT, USA: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.
- [6] S. Mascarenhas and M. Agarwal, A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification, Bengaluru, India: 2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON), 2022.
- [7] N. I. H. A. K. S. M. A. R. a. A. S. U. A. I. Newaz, Adversarial Attacks to Machine Learning-Based Smart Healthcare Systems, Taipei, Taiwan: GLOBECOM 2020 - 2020 IEEE Global Communications Conference, 2020.
- [8] M. a. K. R. Chhabra, An Efficient ResNet-50 based Intelligent Deep Learning Model to Predict Pneumonia from Medical Images, Erode, India: 2022 International Conference on Sustainable Computing and Data Communication Systems, 2022.
- [9] S. G. Ritu Rani, Automated Retinal Disease Classification Using Fine-Tuned ResNet50: A Deep Learning Approach for Early Diagnosis, Bangalore, India: 2025 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), 2025.
- [10] S. Chauhan, Deep Learning-Based Skin type Classification using Fine-Tuned ResNet50 Architecture, Bangalore, India: 2025 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), 2025.
- [11] S. a. N. T. a. L. N. B. a. R. N. S. Pappula, Detection and Classification of Pneumonia Using Deep Learning by the Dense Net-121 Model, Coimbatore, India: 9th International Conference on Advanced Computing and Communication Systems (ICACCS), 2023.
- [12] M. a. O. P. a. O. M. a. M. V. Tshwale, ResNet50 Pretrained Model Based Pneumonia Detection System, Seattle, WA, USA: 2024 IEEE World AI IoT Congress (AIIoT), 2024.

- [13] C.-Y. Lin, B.-H. Lai, H.-F. Ng, W.-Y. Lin and M.-C. Chang, Robust Defense Against Adversarial Attacks with Defensive Preprocessing and Adversarial Training, Las Vegas, NV, USA: 2025 IEEE International Conference on Consumer Electronics (ICCE), 2025.
- [14] A. T. T. T. Ayse Elvan Aydemir, The Effects of JPEG and JPEG2000 Compression on Attacks using Adversarial Examples, Ithaca, NY, United States: Corenell University, 2018.
- [15] M. Z. a. G. A. a. G. D. Hameed, The Best Defense Is a Good Offense: Adversarial Attacks to Avoid Modulation Detection, IEEE Transactions on Information Forensics and Securit, 2021.
- [16] X. a. H. P. a. Z. Q. a. L. X. Yuan, Adversarial Examples: Attacks and Defenses for Deep Learning, IEEE Transactions on Neural Networks and Learning Systems, 2019.
- [17] A. a. K. S. H. a. H. M. a. S. J. a. S. L. Mustafa, Image Super-Resolution as a Defense Against Adversarial Attacks, IEEE Transactions on Neural Networks and Learning Systems, 2020.
- [18] N. a. M. A. a. K. N. a. S. M. Akhtar, Advances in Adversarial Attacks and Defenses in Computer Vision: A Survey, IEEE Access, 2021.
- [19] X. a. H. P. a. Z. Q. a. L. X. Yuan, Adversarial Examples: Attacks and Defenses for Deep Learning, IEEE Transactions on Neural Networks and Learning Systems, 2019.
- [20] A. K. N. P. I. G. D. B. P. M. Florian Tramèr, Ensemble Adversarial Training: Attacks and Defenses, International Conference on Learning Representations, 2017.
- [21] P. S. K. P. B. J. ., S. R. P. H. S. E. Keshav Kansala, Defending against adversarial attacks on Covid-19 classifier: A denoiser-based approach, 2022.
- [22] H. X. H. Z. M. C. a. N. W. S. Cheng, Towards Feature Space Adversarial, Proceedings of the AAAI Conference on Artificial Inrelligence.
- [23] T. Z. a. C. G. X. Yu, Towards Harmonized Regional Style Transfer and Manipulation for Facial Images, Researchgate, 2022.

7. Appendices

Appendix 1 – Key Formulas of the Feature Consistence Defense Technique

Key Formulas for the Detection Mechanism

1. Edge Difference Score

Edge Difference Map: For two edge maps, $E_{\text{orig}}(x, y)$ and $E_{\text{attacked}}(x, y)$, computed using the Canny edge detector, the absolute difference at each pixel is given by:

$$D(x, y) = |E_{\text{orig}}(x, y) - E_{\text{attacked}}(x, y)|$$

Edge Score (Mean Edge Difference): Over N pixels (where N is the total number of pixels), the overall edge score is computed as:

$$\text{Edge Score} = \frac{1}{N} \sum_{x,y} D(x, y)$$

2. Structural Similarity Index (SSIM)

The SSIM index between two grayscale images x and y is defined as:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

where:

- μ_x and μ_y are the mean intensities of images x and y , respectively,
- σ_x^2 and σ_y^2 are the variances,
- σ_{xy} is the covariance between x and y , and
- C_1 and C_2 are small constants for numerical stability.

3. FFT Magnitude

For a grayscale image $I(x, y)$ of size $M \times N$, the 2-D Fast Fourier Transform (FFT) is defined as:

$$F(u, v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} I(x, y) e^{-j2\pi(\frac{xu}{M} + \frac{yv}{N})}$$

The magnitude of the FFT (using a logarithmic scale) is then computed as:

$$\text{Mag}(u, v) = 20 \cdot \log(|F(u, v)| + 1)$$

4. Decision Criteria for Attack Detection

Let:

$$T_{\text{edge}} \quad \text{and} \quad T_{\text{ssim}}$$

be the predefined thresholds for the edge score and SSIM score, respectively.

Then, the detection decision is made as follows:

If Edge Score $> T_{\text{edge}}$ **or** SSIM $< T_{\text{ssim}}$ then, Attack Detected;

otherwise, the image is considered authentic.

Appendix 2 – Algorithm of Style Transfer Manipulation Attack

Algorithm 1 Style Transfer attack Method (STM)

Input: A clean image \mathbf{x} with ground-truth label y , surrogate classifier with parameters θ , and the loss function J .

Parameters: The magnitude of perturbation ϵ ; maximum iteration T ; decay factor μ ; the upper bound of neighborhood β for \mathbf{r} ; the mixing ratio γ ; the number of random generating examples N .

Output: An adversarial example \mathbf{x}^{adv} .

```
1:  $\alpha = \epsilon/T$ ;  
2:  $\mathbf{g}_0 = 0, \mathbf{x}_0^{adv} = \mathbf{x}$ ;  
3: for  $t = 0, 1, \dots, T - 1$  do  
4:   for  $i = 0, 1, \dots, N - 1$  do  
5:     Obtain a random stylized image  $\mathbf{x}_s$  by  $ST(\mathbf{x}_t^{adv}, \mathbf{z})$ ;  
6:     Mix the original image by  $\bar{\mathbf{x}} = \gamma \cdot \mathbf{x} + (1 - \gamma) \cdot \mathbf{x}_s + \mathbf{r}$ ;  
7:     Calculate the gradient  $\bar{\mathbf{g}}_i = \nabla_{\bar{\mathbf{x}}} J(\bar{\mathbf{x}}, y; \theta)$ ;  
8:   end for  
9:   Get the average gradient,  $\bar{\mathbf{g}} = \frac{1}{N} \sum_{i=1}^N \bar{\mathbf{g}}_i$ ;  
10:   $\mathbf{g}_{t+1} = \mu \cdot \mathbf{g}_t + \frac{\bar{\mathbf{g}}}{\|\bar{\mathbf{g}}\|_1}$ ;  
11:  Update  $\mathbf{x}_{t+1}^{adv}$  by  
      
$$\mathbf{x}_{t+1}^{adv} = \text{Clip}_{\mathbf{x}}^{\epsilon} \{ \mathbf{x}_t^{adv} + \alpha \cdot \text{sign}(\mathbf{g}_{t+1}) \};$$
  
12: end for  
13: return  $\mathbf{x}^{adv} = \mathbf{x}_T^{adv}$ .
```

Appendix 3 – Sample Predictions of Baseline ResNet50 Model

True: Pneumonia
Predicted: 1



True: Pneumonia
Predicted: 1



True: Normal
Predicted: 0



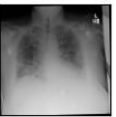
True: Pneumonia
Predicted: 1



True: Pneumothorax
Predicted: 2



True: Pneumonia
Predicted: 1



True: Pneumonia
Predicted: 1



True: Normal
Predicted: 0



True: Pneumothorax
Predicted: 2



True: Pneumonia
Predicted: 1

