

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

B.Sc. (Hons) Degree in Information Technology specialized in Cybersecurity

Department of Computer Systems and Engineering

Sri Lanka Institute of Information Technology

Sri Lanka

April 2025

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

24-25J-079

Final Group Report

Student Details

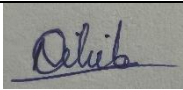



| Student Name | Student ID |
|---------------------|-------------------|
| Premakanthan. N | IT21197550 |
| D.S.C. Wijesuriya | IT21155802 |
| P.G.E.J. Sandamal | IT21166860 |
| W.N. Dilsara | IT21182600 |

Sri Lanka Institute of Information Technology Sri Lanka

April 2025

DECLARATION OF THE CANDIDATE & SUPERVISOR

We declare that this is our own work and this proposal 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 our 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.

| Student Name | Registration Number | Signature |
|-------------------|---------------------|---|
| N. Premakanthan | IT21197550 |  |
| D.S.C. Wijesuriya | IT21155802 |  |
| P.G.E.J. Sandamal | IT21166860 |  |
| W.N. Dilsara | IT21182600 |  |

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

Signature of the supervisor

Date

Abstract

Convolutional Neural Networks (CNNs) have become indispensable in medical image analysis due to their exceptional diagnostic accuracy. However, their vulnerability to adversarial attacks poses a critical challenge, especially in healthcare settings where diagnostic integrity is paramount. This research investigates the resilience of four prominent CNN architectures—ResNet50, DenseNet121, InceptionV3, and VGGNet16—against advanced adversarial attacks, with a focus on the Penalizing Gradient Norm (PGN) method. PGN enhances attack transferability by generating perturbations that guide inputs toward flat local maxima in the loss landscape, increasing the likelihood of misclassification across models. Using chest X-ray images from the NIH dataset, the study simulates attacks including Penalizing Gradient Norm (PGN), Style Transfer Manipulation (STM), Structure Invariant Attack (SIA), and Uniform Scale Mix Mask (USMM) to evaluate model robustness. Defensive strategies such as adversarial training, JPEG compression, high-level representation denoising, and CycleGAN-based restoration were implemented and analyzed for effectiveness. The findings reveal that while PGN poses a significant threat to model performance, targeted defenses—especially adversarial training—can substantially mitigate its impact without degrading diagnostic accuracy, paving the way for safer deployment of AI in medical diagnostics.

Keywords: *Convolutional Neural Networks (CNNs), Medical Imaging, Adversarial Attacks, Style Transfer Manipulation (STM), Structure Invariant Attack (SIA), Uniform Scale Mix Mask ,Penalizing Gradient Norm (PGN)*

TABLE OF CONTENTS

| | |
|--|-----|
| Abstract..... | iv |
| TABLE OF CONTENTS | v |
| TABLE OF FIGURES..... | vi |
| LIST OF ABBREVIATIONS..... | vii |
| 1. Introduction | 1 |
| 1.1 Overview..... | 1 |
| 1.2 Literature Review | 2 |
| 1.3 Research Gap | 3 |
| 1.4 Problem Statement..... | 3 |
| 1.5 Research Objectives..... | 4 |
| 2. Methodology..... | 6 |
| 2.1 Overall System Diagram..... | 6 |
| 2.2 Dataset Selection and Preprocessing..... | 7 |
| 2.3 Selection of Models | 9 |
| 2.4 Implementation of Attacks..... | 11 |
| 2.5 Defense Strategies..... | 15 |
| 2.6 Performance Metrics and Evaluation | 17 |
| 2.7 Tools & Environment | 18 |
| 3. Integration..... | 20 |
| 3.1 System Overview | 21 |
| 4. Commercialization..... | 24 |
| 4.1 Stake Holders | 24 |
| 5. Conclusion..... | 27 |
| 6. References | 28 |

TABLE OF FIGURES

| | |
|---|----|
| Figure 1: Overall System Diagram | 6 |
| Figure 2:Image Counts per Class of Original Dataset | 8 |
| Figure 3: Architecture of ResNet50 Model | 9 |
| Figure 4: Architecture of DenseNet121 Model | 10 |
| Figure 5:Architecture of InceptionV3Model | 10 |
| Figure 6: Architecture of VGGNet-16Model | 11 |
| Figure 7:Implementation of STM Attack | 12 |
| Figure 8: Implementation of PGN Attack..... | 12 |
| Figure 9:Implementation of SIA Attack | 13 |
| Figure 10:Implementation of USMM Attack | 14 |
| Figure 11: System Integration | 21 |

LIST OF ABBREVIATIONS

| Abbreviation | Definition |
|--------------|---|
| CNN | Convolutional Neural Network |
| PGN | Penalizing Gradient Norm |
| SIM | Structure Invariant Attack |
| USSM | Uniform Scale Mix Mask |
| STM | Style Transfer Manipulation |
| IDE | Integrated Development Environment |
| API | Application Programming Interface |
| ML | Machine Learning |
| VGG | Visual Geometry Group (Network) |
| ResNet50 | Residual Network with 50 Layers |
| GPU | Graphics Processing Unit |
| HGD | High-Level Representation Guided Denoiser |
| JPEG | Joint Photographic Experts Group |

1. Introduction

1.1 Overview

Medical imaging is crucial in today's clinical diagnostics, offering non-invasive insights into patient anatomy and pathology essential for informed treatment decisions. In the last ten years, the integration of digital imaging technology and deep learning has transformed the field: Convolutional Neural Networks (CNNs) consistently attain or exceed human-level performance in tasks like pneumonia identification and lung nodule segmentation. Architectures like as ResNet50, DenseNet121, InceptionV3, and VGG16 utilize deep hierarchical feature extraction to identify intricate patterns in chest X-rays, CT scans, and MRIs. Their ability to recognize subtle texture differences and form indicators has facilitated swift, automated analysis at scale, therefore diminishing both diagnostic turnaround time and inter-observer variability.

However, this remarkable potential is mitigated by an increasing amount of evidence indicating that CNNs can be misled by subtle, hostile perturbations. Specifically, style transfer manipulation wherein texture and color data from a non-medical "style" image are included into a diagnostic scan can significantly impair model accuracy without presenting any discernible alteration to the human observer. In addition to style transfer, threats such as structure-invariant perturbations, penalized gradient-norm distortions, and uniform mix-mask masking leverage various elements of a model's dependence on low-level features. These adversarial tactics reveal significant weaknesses: a model attaining over 95% accuracy on unaltered data may experience a performance decline to below 40% when faced with a meticulously designed perturbation.

In acknowledgment of these hazards, our group conducted a thorough assessment of four prominent CNN architectures ResNet50, DenseNet121, InceptionV3, and VGG16—against this varied array of threats. Each network possesses distinct advantages: ResNet's residual connections enable very deep learning; DenseNet's interconnected layers enhance efficient feature reuse; Inception modules acquire multi-scale representations; and VGG's consistent filters offer a clear, interpretable standard. We attempted to assess the relative robustness of each model and identify patterns of shared susceptibility by using style transfer manipulation (STM), structure-invariant assaults (SIA), penalized gradient-norm (PGN) attacks, and uniform mix-mask perturbations.

Importantly, resilience cannot be assessed merely by witnessing failure. We implemented and thoroughly evaluated many defense tactics, examining each method's capacity to recover diagnostic accuracy while preserving critical medical information. Our experiments demonstrated that although adversarial training yields the most significant improvements, alternative defenses frequently result in excessive smoothing or the introduction of artifacts that undermine clinical applicability. These findings highlight a substantial research deficiency: current general-purpose defenses require meticulous adaptation or complete development to satisfy the stringent criteria of medical imaging.

This study delineates our comprehensive methodology, encompassing dataset curation, preprocessing, model training, adversarial attack development, defensive implementation, and performance evaluation. We provide a

comparative review of CNN vulnerabilities together with practical suggestions for incorporating robustification strategies into real-world diagnostic workflows. By highlighting the obstacles and potential opportunities for future endeavors, we seek to provide AI developers, medical software suppliers, and healthcare institutions with the necessary information to implement reliable and secure deep learning systems in the vital field of medical imaging.

1.2 Literature Review

The Style Transfer Manipulation (STM) technique, based on style transfer methods initially presented by Gatys, facilitates the modification of visual styles while maintaining content integrity. Subsequent research, including that of Zhang, illustrated the effectiveness of STM in producing adversarial cases for scene text recognition algorithms, exposing considerable vulnerabilities [1]. Nevertheless, the utilization of STM in medical imaging is still inadequately investigated, especially with the use of style-based attacks to assess model resilience. Huang's investigation of regional style transfer in medical imaging highlights the potential of these technologies but fails to consider their antagonistic consequences [2].

Structure Invariant Attack (SIA) is a revolutionary technique for improving the transferability of adversarial examples. Wang propose that SIA segments input images into blocks and implements localized modifications while maintaining the global structure [3]. This method markedly enhances transferability among various models, including CNNs and transformers. Notwithstanding its achievements in other fields, the implications of SIA for medical imaging have not been sufficiently examined, especially regarding diagnostic interruptions and the robustness of defense mechanisms.

The Penalizing Gradient Norm (PGN) assault, as presented in recent studies, attains enhanced transferability by directing adversarial samples towards flat local maxima. By imposing a penalty on the gradient norm during optimization, PGN guarantees that adversarial examples maintain robustness across diverse models. Ge confirmed that adversarial cases situated in flat regions demonstrate enhanced transferability. Although PGN has been thoroughly examined in general-purpose contexts, its utilization in medical imaging and efficacy against current defense mechanisms are still inadequately investigated [4].

The Uniform Scale Mix Mask (USMM) technique, presented by Wang, improves the transferability of adversarial instances through uniform scaling and blending of image segments. While preliminary studies concentrated on general-purpose datasets, the technique's applicability to medical imaging is becoming increasingly apparent. Research conducted by Rana and Sharma emphasizes the exceptional accuracy of convolutional neural networks (CNNs) in medical diagnostics, although neglects to include the influence of advanced attacks such as universal adversarial perturbations (USMM) [5]. Existing protection mechanisms exhibit insufficient specificity against USMM attacks, resulting in a significant vulnerability in the robustness of medical imaging models.

The research conducted by Arjun Thangaraju and Cory Merkel provides a comprehensive analysis of adversarial attacks and defenses in deep learning, focusing on image categorization. Prior to examining various assault strategies and corresponding response techniques, it delineates hostile examples and elucidates their significance [6]. This study use the Fast Gradient Signed Method (FGSM) to demonstrate how adversarial perturbations can substantially diminish the accuracy of a CNN classifier on the MNIST dataset. The authors illustrate how adversarial training effectively mitigates these threats, enhancing the model's robustness and restoring its accuracy.

1.3 Research Gap

Although Convolutional Neural Networks (CNNs) have advanced considerably in medical imaging, especially in diagnosing ailments from chest X-rays, they are still susceptible to adversarial attacks—subtle, meticulously designed alterations that can mislead models into erroneous predictions. Within the expansive domain of computer vision, various adversarial attacks, including Style Transfer Manipulation (STM), Penalizing Gradient Norm (PGN), Structure Invariant Attack (SIA), and Uniform Scale Mix Mask (USMM), have been meticulously developed and evaluated on natural images, such as those from datasets like CIFAR-10, ImageNet, and COCO. These assaults have revealed the capacity to markedly diminish model efficacy, undermine current protections, and reveal inherent flaws in CNN systems.

A significant research gap persists in the use and evaluation of these sophisticated adversarial strategies inside the medical imaging field, particularly concerning chest X-rays. In contrast to natural images, medical images exhibit distinctive structural, anatomical, and semantic attributes crucial for precise diagnosis. Attacks formulated for general image datasets may exhibit divergent or even more detrimental effects when applied to medical images, owing to the heightened sensitivity of clinical models to minor changes in diseased areas.

1.4 Problem Statement

Despite the crucial role of CNNs in illness identification from chest X-rays, their susceptibility to adversarial attacks poses a significant worry that remains largely unexamined within the realm of medical imaging. The majority of current adversarial assault research has focused on broad picture datasets and tasks, with minimal adaptation or validation in medical contexts.

Advanced adversarial strategies, including Style Transfer Manipulation (STM), Penalizing Gradient Norm (PGN), Structure Invariant Attack (SIA), and Uniform Scale Mix Mask (USMM), have demonstrated considerable efficacy in deceiving convolutional neural networks (CNNs) trained on genuine pictures. Nonetheless, their efficacy, conduct, and possible ramifications in the context of chest X-ray diagnosis remain uncertain. The absence of inquiry into domain-specific adversarial threats engenders a significant gap in the existing corpus of knowledge, particularly given the life-critical implications of medical predictions.

1.5 Research Objectives

1. Implementation of Adversarial Attacks

This study seeks to execute and evaluate the impact of sophisticated adversarial attack methodologies on convolutional neural network models employed for chest X-ray image classification. The research will concentrate on four notable attack techniques: Style Transfer Manipulation (STM), Penalizing Gradient Norm (PGN), Structure Invariant Attack (SIA), and Uniform Scale Mix Mask (USMM). While these attacks have been thoroughly examined in relation to natural picture datasets, they have not been fully investigated within the realm of medical imaging. This work will assess the impact of attacks on model decision-making in clinical contexts, given that chest X-rays possess essential diagnostic characteristics that are frequently subtle and spatially concentrated. This purpose encompasses comprehending the characteristics of the induced perturbations, their visual inconspicuousness, and their capacity to mislead diagnostic models, potentially leading to detrimental misclassifications.

2. Evaluate the Robustness of the ResNet CNN Model

This objective is to evaluate the resilience of the ResNet CNN architecture against adversarial attacks. This project will assess the effectiveness of a well-trained ResNet model for chest X-ray categorization against various types of adversarial perturbations. Key performance measures, including accuracy, precision, recall, and F1-score, will be evaluated on both pristine and hostile datasets. This assessment will underscore the extent to which the attacks impair the model's capacity to accurately diagnose diseases and will determine if specific classes or categories of anomalies exhibit greater susceptibility. The study seeks to measure the degree of performance deterioration and understand model behavior in adversarial conditions.

3. Test and Validate Existing Defense Mechanisms

The project will investigate and test current protection mechanisms to enhance the resilience of CNNs in response to recognized vulnerabilities. Specifically, methodologies such as data augmentation and adversarial training will be incorporated into the training pipeline of the ResNet model. Data augmentation applies diverse changes to the training dataset, promoting improved generalization of the model, whereas adversarial training incorporates adversarial samples during training to bolster resilience to analogous attacks. The aim is to evaluate the efficacy of these defenses in mitigating or diminishing the effects of hostile inputs, especially in a medical imaging context where even little model inaccuracies can lead to significant repercussions.

4. Measure the Effectiveness of Defense Strategies

This objective entails a comprehensive assessment of the effectiveness of the established defense techniques in bolstering the model's resilience to adversarial attacks. The analysis will evaluate model performance prior to and subsequent to the implementation of defense mechanisms, emphasizing enhancements in classification metrics and decreases in assault success rates. The study will evaluate whether the defenses preserve performance on clean, unaltered data, ensuring that enhanced robustness does not compromise diagnostic accuracy. This stage is essential for confirming the practical applicability of the defenses in actual medical contexts and for aiding in the advancement of dependable AI systems in clinical practice.

2. Methodology

2.1 Overall System Diagram

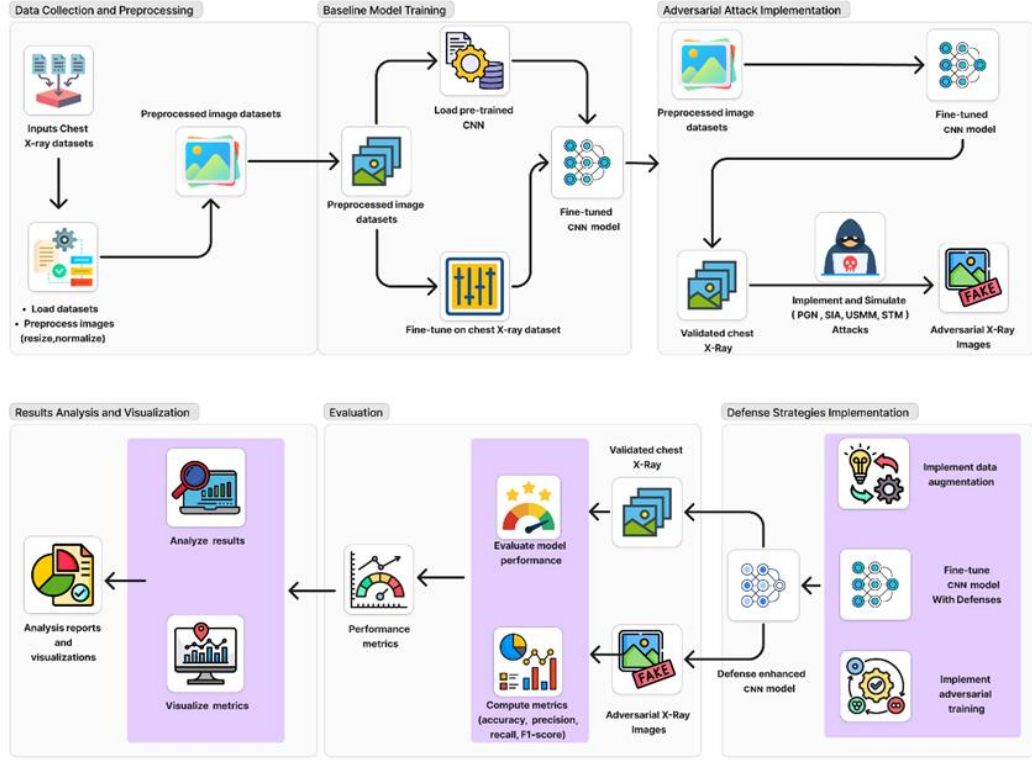


Figure 1: Overall System Diagram

The workflow begins with the Data Collection and Preprocessing stage. Chest X-ray datasets, such as the NIH Chest X-ray dataset, are collected and preprocessed to prepare them for model training. The preprocessing includes resizing the images to 224x224 pixels and normalizing the pixel values to improve model convergence and consistency. Additionally, data augmentation is applied by rotating images by ± 10 degrees to introduce variability and enhance model generalization. This results in a preprocessed dataset ready for training the CNN models.

In the Baseline Model Training stage, pre-trained CNN architectures such as DenseNet121, ResNet50, VGGNet19, and InceptionV3 are loaded. These models, initially trained on large datasets like ImageNet, are fine-tuned on the chest X-ray dataset to adapt them to the medical imaging domain. Fine-tuning involves unfreezing layers and adding additional layers to improve performance on the specific task of chest X-ray classification. This step produces finetuned CNN models capable of classifying clean (unaltered) medical images.

The Adversarial Attack Implementation stage focuses on evaluating the robustness of the trained CNN models by simulating adversarial attacks. The preprocessed chest X-ray images are subjected to adversarial perturbations using techniques such as Style Transfer Manipulation (STM), Penalizing Gradient Norm (PGN), Structure Invariant Attack (SIA), and Uniform Scale Mix Mask (USMM). These attacks create adversarial images that appear visually similar to the original images but are designed to mislead the CNN models into making incorrect predictions. This step results in adversarial X-ray images that test the vulnerability of the models.

Next, the Defense Strategies Implementation stage aims to enhance the resilience of the CNN models against these adversarial attacks. Defensive techniques such as data augmentation, adversarial training, and randomized smoothing are applied. Data augmentation involves introducing rotated images into the training process, while adversarial training involves retraining the models using both clean and adversarial examples to improve robustness. Randomized smoothing adds noise to the input during inference to mitigate the impact of adversarial perturbations. The fine-tuned models are re-trained with these defenses, resulting in defense-enhanced CNN models that are more resilient to adversarial attacks.

In the Evaluation stage, the performance of the defense-enhanced models is assessed using key metrics such as accuracy, precision, recall, and F1-score. The models are evaluated on both clean chest X-ray images and adversarial X-ray images to determine how well the defense mechanisms protect against adversarial threats. This evaluation helps quantify the effectiveness of the defenses.

Finally, the Results Analysis and Visualization stage involves analyzing and visualizing the results of the evaluation. The performance metrics are analyzed, and the results are presented through charts, graphs, and other visual tools. These visualizations help illustrate the impact of adversarial attacks on the models and the success of the implemented defense strategies. The final output includes comprehensive analysis reports and visual representations, providing insights into the robustness and security of CNN-based medical imaging systems. This end-to-end workflow ensures a systematic approach to assessing the robustness of medical imaging models, implementing adversarial threats, applying defensive measures.

2.2 Dataset Selection and Preprocessing

This research utilized the NIH Chest X-ray Dataset as the primary dataset owing to its comprehensive array of labeled medical pictures and its recognized reliability within the medical imaging research community. The dataset contains around 112,000 frontal-view chest X-ray images labeled with 14 distinct thoracic disease classifications, including pneumonia, pneumothorax, cardiomegaly, among others. The dataset's diversity, scale, and high resolution render it an optimal choice for training and assessing deep learning models for automated disease identification.

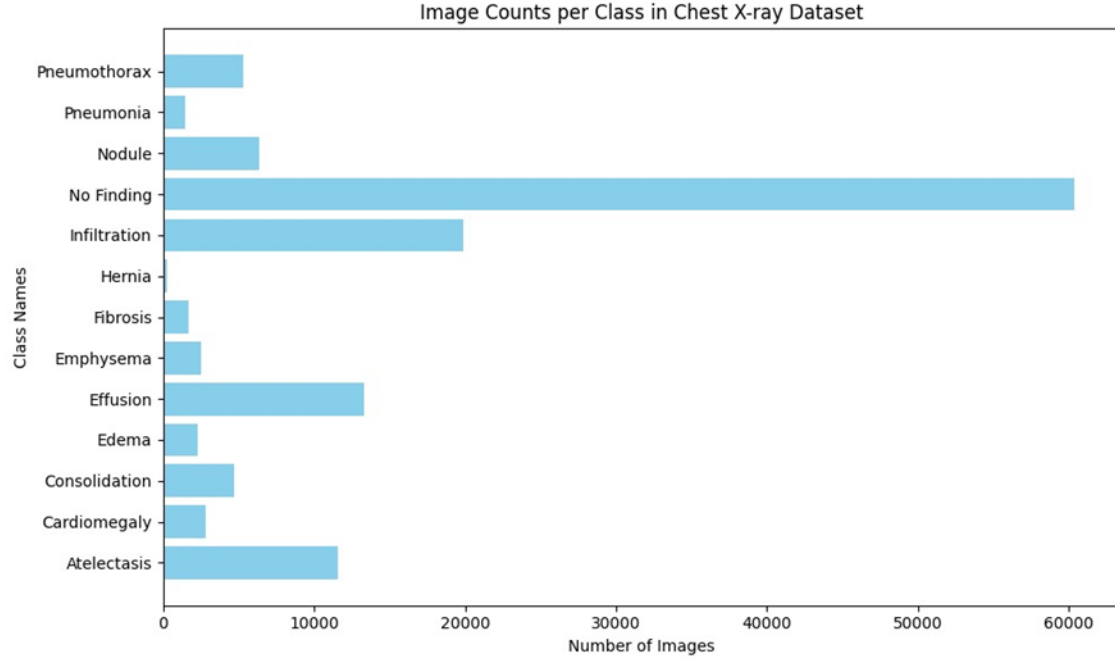


Figure 2: Image Counts per Class of Original Dataset

This study aims to evaluate the influence of adversarial attacks on clinically important illnesses by selecting three categories from the dataset: Normal (no findings), Pneumonia, and Pneumothorax. The categories were selected due to their clinical relevance, prevalence in diagnostic practice, and visible differentiation in chest radiography, offering a substantial background for evaluating model resilience in hostile scenarios.

A balanced subset of 6,000 images per class was compiled for each of the three classes to maintain consistency in training and evaluation. In cases where the dataset lacked a complete set of 6,000 original photos for a specific class—especially for classes with limited examples as pneumothorax—a straightforward yet efficient data augmentation technique was utilized. Image rotation of 5 degrees was specifically performed to the current samples. This method enabled the dataset to be standardized across categories without compromising the semantic integrity of the medical features in the X-rays. The minor rotation maintained the diagnostic integrity while enhancing dataset variability, hence improving model generalization and assuring equitable representation among the chosen classes.

The dataset was standardized by equalizing the number of samples in each class and used minimum augmentation, facilitating robust and equitable comparisons during the training, testing, and adversarial assessment of the CNN models. The meticulous curation and preparation phase was crucial for generating a uniform and therapeutically relevant dataset that corresponds with the goals of adversarial robustness assessment in medical picture classification.

2.3 Selection of Models

This study selected four advanced Convolutional Neural Network (CNN) architectures—ResNet50, DenseNet121, InceptionV3, and VGGNet16—for the categorization of chest X-ray pictures and the assessment of adversarial robustness. These models are extensively utilized in image classification tasks and have exhibited robust performance across numerous medical imaging difficulties. Each model has a distinct architectural design and learning methodology that variably influences feature extraction, generalization, and resilience to adversarial perturbations.

1. ResNet50

ResNet50 is a 50-layer deep convolutional neural network that employs residual learning, a method that mitigates the vanishing gradient issue in deep designs. The primary novelty of ResNet is the residual block, which facilitates the direct flow of gradients using identity shortcuts, circumventing multiple convolutional layers. This facilitates the development of far deeper networks without a decline in performance. ResNet50 is recognized for its proficiency in learning intricate characteristics and has demonstrated significant efficacy in medical imaging applications owing to its capacity to collect both high- and low-level data. Its profundity and structural integrity render it especially adept at discerning nuanced patterns in chest X-rays, including initial indications of pneumonia or pneumothorax.

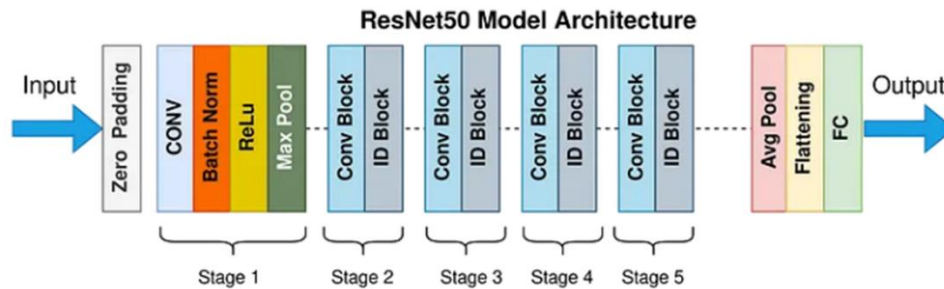


Figure 3: Architecture of ResNet50 Model

2. DenseNet121

DenseNet121 enhances information flow and feature reuse through the implementation of dense connectivity. In this architecture, each layer obtains inputs from all prior layers and transmits its feature maps to all following layers. This design markedly diminishes the issue of redundant features and promotes feature propagation over the network. DenseNet121 is computationally cheap and typically necessitates fewer parameters than other models of same depth, rendering it appealing for medical applications where overfitting is a concern. Its capacity to reuse features renders it especially proficient at identifying subtle distinctions in chest X-ray pictures, which is essential for differentiating between apparently analogous illnesses.

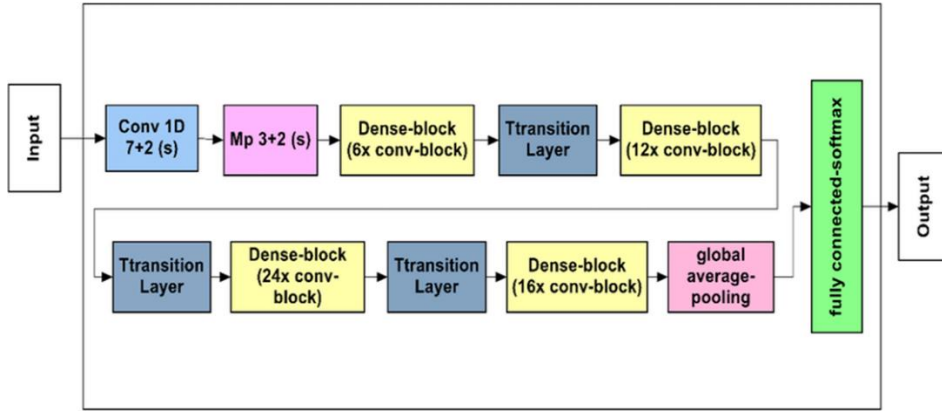


Figure 4: Architecture of DenseNet121 Model

3. InceptionV3

InceptionV3 is an enhanced iteration of the original GoogLeNet design that employs Inception modules specialized components enabling the network to execute several convolution operations (1x1, 3x3, 5x5) concurrently within a single layer. This approach allows the model to simultaneously capture information across several scales and levels of abstraction. InceptionV3 employs methodologies including factorized convolutions, label smoothing, and auxiliary classifiers to enhance training efficiency and precision. The multi-scale feature extraction capacity is particularly advantageous in chest X-ray analysis, where anomalies can differ markedly in size, shape, and location.

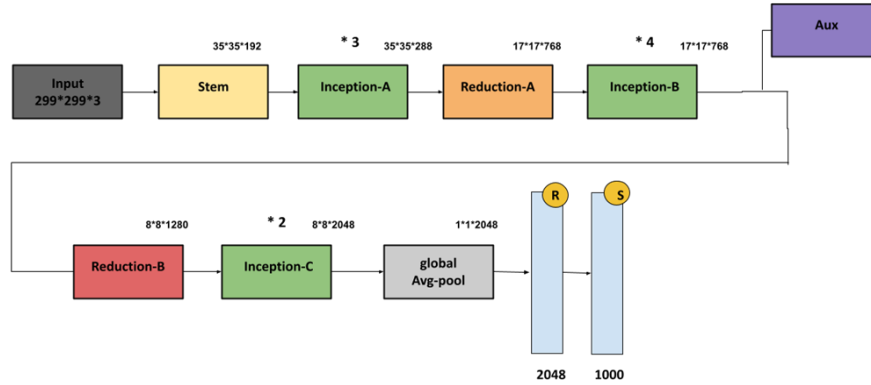


Figure 5: Architecture of InceptionV3 Model

4. VGGNet-16

VGGNet16 is a deep neural network with 16 weight layers, distinguished by its straightforward and consistent architecture, employing exclusively 3x3 convolutional filters and 2x2 max-pooling layers throughout. Notwithstanding its simplicity, VGGNet16 is formidable and has demonstrated robust performance in numerous image recognition tasks. Its efficacy is attributed to its depth and consistency, allowing it to comprehend intricate spatial hierarchies within images. VGGNet16 is frequently employed as a baseline model in medical imaging because of its straightforward installation and interpretability. Nonetheless, its considerable number of factors

may result in increased computational requirements, necessitating a balance between accuracy and generalization efficacy.

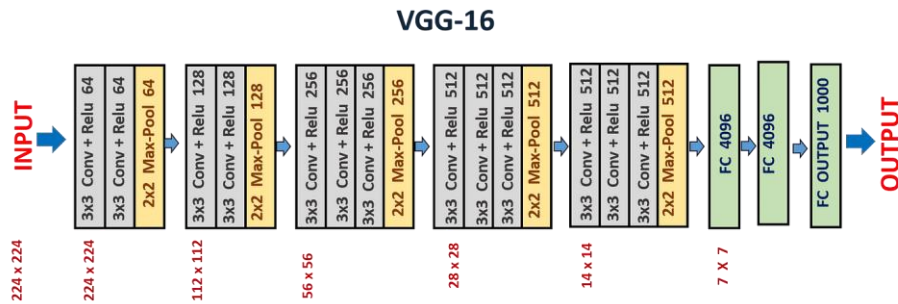


Figure 6: Architecture of VGGNet-16Model

2.4 Implementation of Attacks

1. Style Transfer Manipulation Attack

This study employs style transfer attacks through a sophisticated method that integrates stylistic elements from an external, non-medical design image with the inherent content of a medical image. A distinctive design image is meticulously selected due to its rare color patterns and textures, which are atypical in conventional medical imaging. This decision is essential as it introduces distinctive stylistic features that might subtly yet effectively modify the visual representation of the target medical image. The style elements to be integrated with the original image derive from the chosen design image.

The assault employs a neural style transfer method to amalgamate the two images subsequent to the selection of the style image. This method extracts the content representation of the medical image while simultaneously capturing the decorative elements, such as texture and color, of the design image. Style loss and content loss represent two opposing objectives that are reconciled through the optimization of a composite loss function to attain fusion. The content loss component ensures the preservation of critical diagnostic features and structural integrity of the original medical image, while the style loss component accelerates the incorporation of novel textural and color aspects from the design image.

The program use an iterative gradient descent method to achieve this equilibrium. The pixel values of the original image are incrementally altered in each iteration to progressively incorporate the stylistic elements of the design image. This procedure continues until the generated adversarial image, referred to as the STM (Style Transfer Manipulation) assaulted image, exhibits significant stylistic alterations while remaining strikingly similar to the original. These minor alterations suffice to deceive CNN algorithms into misclassifying the adversarial image, which seems nearly indistinguishable to human observers.

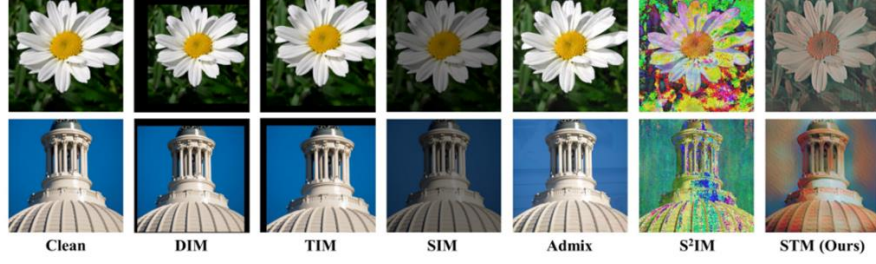


Figure 7: Implementation of STM Attack

2. Penalizing Gradient Norm

The Penalizing Gradient Norm (PGN) attack is designed to improve adversarial transferability by generating adversarial examples that reside within flat local regions of the neural network’s loss landscape. The core hypothesis of this method is based on the empirical observation that adversarial examples in flatter local maxima—where gradients of the loss function change gradually—are more likely to successfully transfer to different models, compared to examples in sharp, sensitive regions. PGN leverages this insight by introducing a penalty on the gradient norm, which effectively pushes adversarial examples toward these stable, flat areas.

Directly computing and optimizing gradients and Hessians to identify these flat regions can be computationally challenging. Therefore, PGN employs an efficient approximation technique called the finite difference method. Instead of computing the full Hessian matrix, it approximates Hessian-vector products by interpolating gradients from closely sampled points within the neighborhood of the current adversarial example. Additionally, PGN reduces variability introduced by random sampling by averaging the gradients of several randomly selected neighboring points, resulting in more stable and reliable gradient estimates.

Extensive experiments and visualizations conducted on standard datasets demonstrate the effectiveness of PGN. Adversarial examples generated by PGN consistently exhibit significantly improved transferability, effectively deceiving multiple models beyond the original target. Furthermore, visual analyses confirm that PGN adversaries occupy smoother, flatter regions of the loss landscape, validating the underlying hypothesis that flat local maxima are beneficial for transfer-based attacks. This method can be seamlessly combined with existing adversarial techniques, thereby further enhancing its practicality and applicability in real-world scenarios.

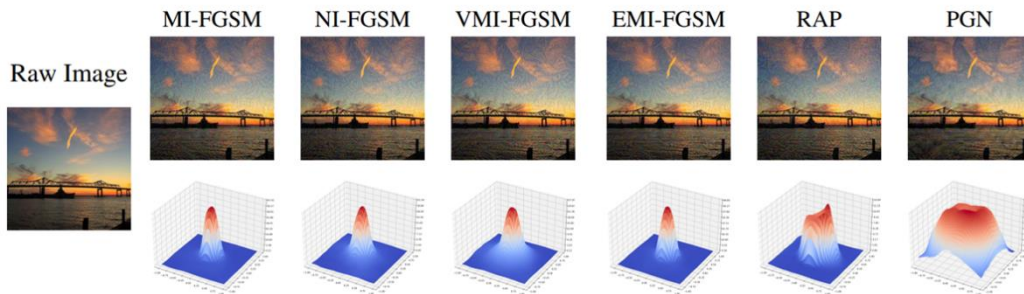


Figure 8: Implementation of PGN Attack

3. Structure Invariant Attack

By altering visual style components while maintaining the image's underlying structural content, the Structure Invariant Attack (SIA) is a new class of adversarial assaults that target deep learning models, especially Convolutional Neural Networks (CNNs). In delicate domains like medical imaging, where clinical judgments are relied on an image's visual information, this method is particularly risky. SIAs concentrate on changing an image's style, such as color tones, texture, or contrast, in a way that is aesthetically pleasing or largely undetectable to human observers, but that severely impairs model performance, in contrast to traditional adversarial attacks that introduce pixel-level noise to trick models.

The fact that SIAs circumvent established protection measures is one of the main reasons they are so dishonest. A lot of common adversarial defenses are designed to pick up on high-frequency noise or odd changes at the pixel level. SIA-modified photos frequently avoid these protections unnoticed since they just change the style and keep the low-level architecture. This reveals serious flaws in CNN models that were only trained on static datasets with limited stylistic diversity. Furthermore, the stakes are enormous in medical imaging systems. Incorrect diagnoses or postponed treatment may result from a little misclassification in a chest MRI or X-ray brought on by a SIA. This possibility of practical consequences emphasizes how urgent it is to research, identify, and protect against SIA dangers.

The fact that SIA attacks are content-preserving, which means they may be executed without access to internal architecture or model gradients, is another issue. This puts SIAs in line with black-box attacks, which makes them more useful and scalable in real-world situations, particularly for attackers who don't know much about the system.

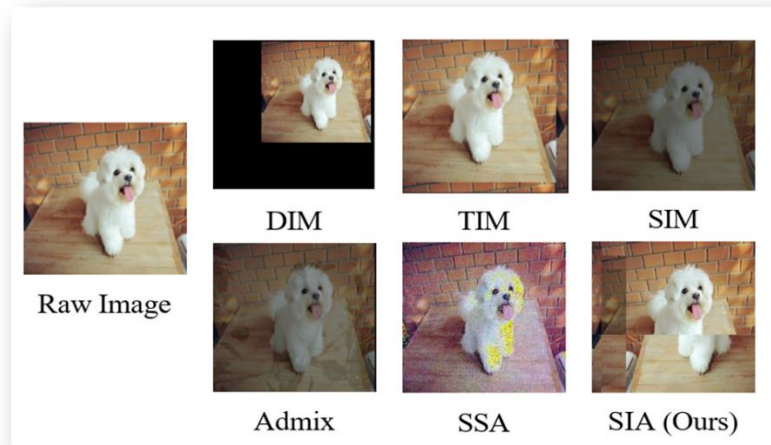


Figure 9: Implementation of SIA Attack

4. Uniform Scale and Mix Mask (USMM) assault

The Uniform Scale and Mix Mask (USMM) attack is a potent and covert adversarial technique that improves the transferability of adversarial examples through the integration of two innovative strategies: Uniform Scale (US) and Mix Mask (MM) transformations. In contrast to conventional attacks that depend on fixed-scale inputs or

basic pixel-level noise, USMM functions in both transformation-space and feature-space, rendering it far more effective and challenging to counter. The Uniform Scale component samples continuous scaling factors from a specified range to produce many scaled iterations of the input image. This technique reveals vulnerabilities in models without scale invariance, as objects of atypical sizes can disrupt the model's learnt representations and result in misclassification. By utilizing many scaled photos during the assault, USMM adeptly captures multi-scale features, enabling it to circumvent models that depend significantly on size consistency in object detection.

The second component, Mix Mask, surpasses previous mixup-based methodologies (such as Admix) by utilizing a multiplicative mask obtained from a picture of a different class, instead of a linear pixel combination. This mask selectively enhances or attenuates areas of the source image according to the configuration of the foreign item. This method discreetly integrates characteristics from a different class into the original image without compromising its overall aesthetic, in contrast to direct pixel substitution or additive noise. The altered image preserves its visual identity yet elicits erroneous internal activations within the model, resulting in misclassification of the image. Element-wise multiplication enables perturbations to remain constrained and bidirectional, softly augmenting or diminishing pixel brightness in a deliberate fashion.

The efficacy of USMM lies in its capacity to evade detection by conventional preprocessing defenses such as JPEG compression or denoising filters, which generally address additive perturbations. USMM alters images via geometric changes and organized feature amalgamation instead of pixel-level noise, resulting in adversarial examples that are aesthetically plausible and semantically coherent, yet deceive the model's internal representations. The assault has been empirically demonstrated to surpass prior state-of-the-art attacks, including SIM and Admix, across multiple CNN architectures. This illustrates USMM's significant transferability and its capacity to generalize across many model types. To effectively counter such an assault, architectural enhancements are necessary to foster intrinsic resilience to scale fluctuations and concealed feature interference, rather than depending exclusively on input sanitization or noise-based detection techniques.

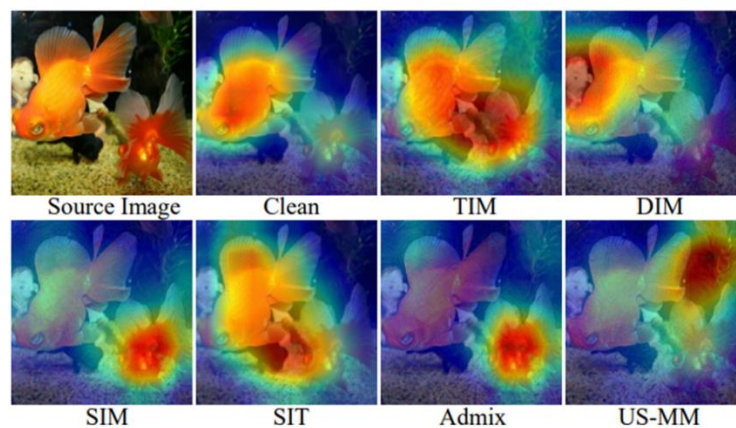


Figure 10: Implementation of USMM Attack

2.5 Defense Strategies

1. Adversarial Training

Adversarial training enhances the conventional training process by including adversarially modified instances into the training dataset. Throughout each epoch, the model encounters both pristine chest X-ray images and their style-transferred or otherwise altered versions. Adversarial training compels the network to learn from both genuine and adversarial inputs, promoting the model's development of feature representations that are invariant to such perturbations. This entails the real-time generation of adversarial samples utilizing the identical assault algorithm under evaluation, subsequently integrating them with the clean batch during loss computation. Over time, the model's decision boundaries adjust to accurately categorize images despite minor, targeted distortions, therefore significantly enhancing robustness against the specific threats encountered during training.

2. High-Level Representation Guided Denoiser (HGD)

The High-Level Representation Guided Denoiser (HGD) is a defensive strategy that mitigates adversarial noise at the feature level instead of processing pixel values directly. In HGD, an independent denoising network is trained using a loss function that penalizes discrepancies between the target classifier's high-level activations on pristine images and those on denoised images. Specifically, with a pristine image and its adversarially modified counterpart, the denoiser is refined to ensure that, upon processing both images through the pre-trained CNN, their internal feature maps (from a selected layer) are as similar as possible. This method mitigates the "error amplification" phenomenon associated with conventional pixel-based denoisers and has demonstrated efficacy in generalizing across both white-box and black-box attacks. In our trials, HGD shown diminished efficacy on chest X-rays owing to the nuanced characteristics of X-ray texture, resulting in excessive smoothing and the loss of essential diagnostic information.

3. JPEG Compression

JPEG compression utilizes the lossy quantization characteristic of the JPEG technique to eliminate high-frequency components, which frequently contain adversarial perturbations. Saving and reloading the adversarial image with modest compression quality (e.g., 70–90%) might substantially diminish the tiny pixel-level noise generated by style transfer or gradient-based attacks. This preprocessing step is simple to do and entails negligible computational burden. Nonetheless, when used for chest X-rays—where nuanced gradients and minor textural differences are critical for diagnosis—JPEG compression can compromise authentic image details in conjunction with adversarial noise, occasionally diminishing total model efficacy more than it enhances resilience.

4. CycleGAN-Based Restoration

CycleGAN-Based Restoration CycleGAN is an unpaired image-to-image translation network that establishes mappings between two domains (namely, "adversarial" and "clean" X-rays) through cycle consistency. Two generators are developed: one to transform hostile images into pristine X-rays and another to execute the inverse process. Domain-specific discriminators uphold realism, but cycle-consistency losses guarantee that a round-trip translation reproduces the original image. This theoretically enables the elimination of stylistic distortions without the need for matched samples. In practice, our CycleGAN models frequently generated artifacts or inadequately removed style-transfer noise, suggesting that unpaired translation may have difficulty maintaining intricate medical structures.

5. Feature Consistency Defense for STM Attack

Feature Consistency Defense for STM Attack is the repetitive modification of an input image to align with two distinct sets of target representations: the high-level content features of an untainted image and the style statistics (such as Gram matrices) of a reliable reference. The approach employs a pretrained network, such as a VGG backbone, to calculate content loss (L_2 distance between feature maps) and style loss (L_2 distance between Gram matrices), thereafter executing gradient-based optimization on the input pixels. The resultant image seeks to "rectify" hostile perturbations by reverting them to the authentic content-style manifold. This optimization, although theoretically promising, can be sluggish and may cause distortions if the equilibrium between content and style terms is not meticulously calibrated, hence complicating real-time implementation in a clinical environment.

6. Hybrid Defense Against SIA Attacks

The hybrid defense integrates moderate JPEG compression (quality 50–80%) with a subsequent Total Variation Minimization (TVM) step to mitigate structure-invariant attack distortions while maintaining essential diagnostic data. The lossy encoding of JPEG effectively eliminates high-frequency noise and minor texture distortions caused by the attack, while TVM subsequently smooths any residual spatial inconsistencies without compromising edges or anatomical features. Adjusting the TVM smoothing parameter to a modest threshold guarantees that only residual artifacts are diminished, yielding input images that nearly mirror the pristine X-rays. The simultaneous application of these two preprocessing processes markedly enhances model accuracy against SIA attacks while preserving the clinical quality of the images.

2.6 Performance Metrics and Evaluation

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

Accuracy quantifies the entire ratio of right predictions—comprising both true positives (TP) and true negatives (TN)—relative to the total number of cases. In medical imaging, it offers a rapid, overarching perspective on the frequency with which the CNN accurately assigns diagnoses; yet, it may be deceptive when instances of disease are far uncommon than normal cases. During adversarial attacks, a significant decline in accuracy promptly indicates that the model's global decision boundary has been undermined, regardless of our inability to ascertain whether it is overlooking genuine illnesses or fabricating erroneous ones.

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

Precision measures the dependability of positive predictions by calculating the ratio of genuine disease cases to all photos identified as diseased. High precision signifies that when the model generates an alert—such as signaling pneumonia—it is improbable to be a false positive, therefore minimizing unnecessary follow-up testing and alleviating patient worry. In response to adversarial perturbations, a decrease in precision indicates that the model is excessively responsive to little noise, misinterpreting innocuous patterns as pathological.

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

Recall (or sensitivity) quantifies the model's ability to identify all actual disease instances, assessing the proportion of true positives identified among all genuine cases. In clinical practice, strong recall is essential, as failing to identify an actual pathology (a false negative) might postpone treatment and jeopardize patient outcomes. A decline in recall due to adversarial manipulation signifies that the attack is successfully obscuring essential diagnostic traits, leading the model to neglect authentic disease indicators.

$$\text{F1 Score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

The F1 score integrates precision and recall into a unified metric, balancing the model's necessity to minimize both false positives and false negatives. It is particularly beneficial in imbalanced or antagonistic contexts, when an enhancement in one metric may detrimentally affect another. Monitoring the F1 score enables us to evaluate if defensive tactics enhance the model's overall discriminative capability, rather than merely balancing sensitivity and specificity.

2.7 Tools & Environment

1. Programming Language

All tests and implementations were executed in Python (v3.8+), selected for its clarity and comprehensive array of scientific libraries. The straightforward syntax of Python enabled swift prototyping of intricate attack and defense strategies, while its extensive community support provided access to well-maintained, efficient tools for data management, model creation, and assessment.

2. Deep Learning Frameworks

TensorFlow and Keras: Employed for the development and training of the foundational CNN architectures (ResNet50, DenseNet121, InceptionV3, VGG16). Keras's high-level API facilitated rapid model development and iteration, whilst TensorFlow's backend offered efficient GPU acceleration, distributed training capabilities, and seamless model serialization in the SavedModel format for subsequent inference.

PyTorch: Utilized for the creation of bespoke adversarial and denoising modules. The dynamic computing network facilitated real-time gradient analysis and debugging, essential for optimizing iterative style-transfer processes, Total Variation Minimization, and CycleGAN training cycles.

3. Data Management and Preprocessing

Pandas: Administered picture metadata (file locations, labels, split assignments) using DataFrames, optimizing the creation of Keras and PyTorch data loaders.

NumPy: Functioned as the cornerstone for all numerical operations, encompassing batch-level transformations, bespoke tensor manipulations for attack development, and efficient in-memory computations of performance metrics.

4. Visualization and Evaluation

Matplotlib: Utilized to illustrate training and validation accuracy and loss curves, elucidating convergence behavior prior to and subsequent to unfreezing layers and implementing defense strategies.

Seaborn: Produced sophisticated statistical visualizations—confusion matrices and distribution graphs illustrating metric variations under attack—to effectively convey model performance outcomes.

Scikit-learn: Comprehensive implementations offered is for calculating accuracy, precision, recall, and F1-score, along with tools for data partitioning, classification reporting, and ROC curve analysis as required.

5. Development Environment and Collaboration

Google Colab: Functioned as the principal integrated development environment for swift experimentation, including complimentary access to NVIDIA Tesla GPUs. Collaborative notebooks enabled instantaneous cooperation among team members, permitting peer evaluation of model structures and attack implementations.

6. Hardware and Cloud Infrastructure

Local Workstations: Employed for lightweight testing and debugging on CPU or single-GPU configurations.

Google Cloud Platform (GCP): Implemented larger virtual machine instances with advanced GPUs (Tesla V100/P100) and extensive disk storage for large-scale attack batch generation and adversarial training. The scalability of GCP facilitated a smooth shift from prototyping on Colab to comprehensive training executions without necessitating code alterations.

3. Integration

Ensuring strong cybersecurity measures has grown more and more important in the quickly changing digital landscape, especially in delicate industries like healthcare. Cyber threats pose a serious threat to healthcare systems today, especially hostile assaults that compromise the accuracy and integrity of patient records and medical diagnostic instruments. Understanding these difficulties, our group has created a cutting-edge integrated solution that is intended to protect healthcare applications from these dangers.

This outlines a thorough integration strategy that systematically integrates four advanced models specifically designed to identify, evaluate, and effectively counteract adversarial threats. By addressing distinct facets of cybersecurity, these models improve the overall security posture and robustness of healthcare systems. This system improves operational reliability, user trust, and compliance with healthcare data rules in addition to ensuring data safety using cutting-edge technologies and meticulous integration procedures.

In addition to secure and scalable data management solutions, our integration strategy entails careful coordination between contemporary frontend and backend technologies. High-performance online experiences, seamless user interactions, and quick rendering are guaranteed when NextJS is used for the frontend. Python-powered Flask effectively handles computationally demanding activities on the backend, enables real-time data processing, and enables complex machine learning processes that are necessary for adversarial detection. Furthermore, Google Drive's safe cloud storage features are integrated into our system, facilitating easy data management and accessibility.

Vendors, developers of healthcare solutions, and system administrators are just a few of the stakeholders whose varied needs are clearly addressed in the integration strategy. Ease of use, strong security, and effective administration are guaranteed by the system's meticulous alignment of its functionalities with stakeholders' operating requirements. The resulting integrated solution protects patient data and organizational integrity by offering complete defense against advanced cyberthreats.

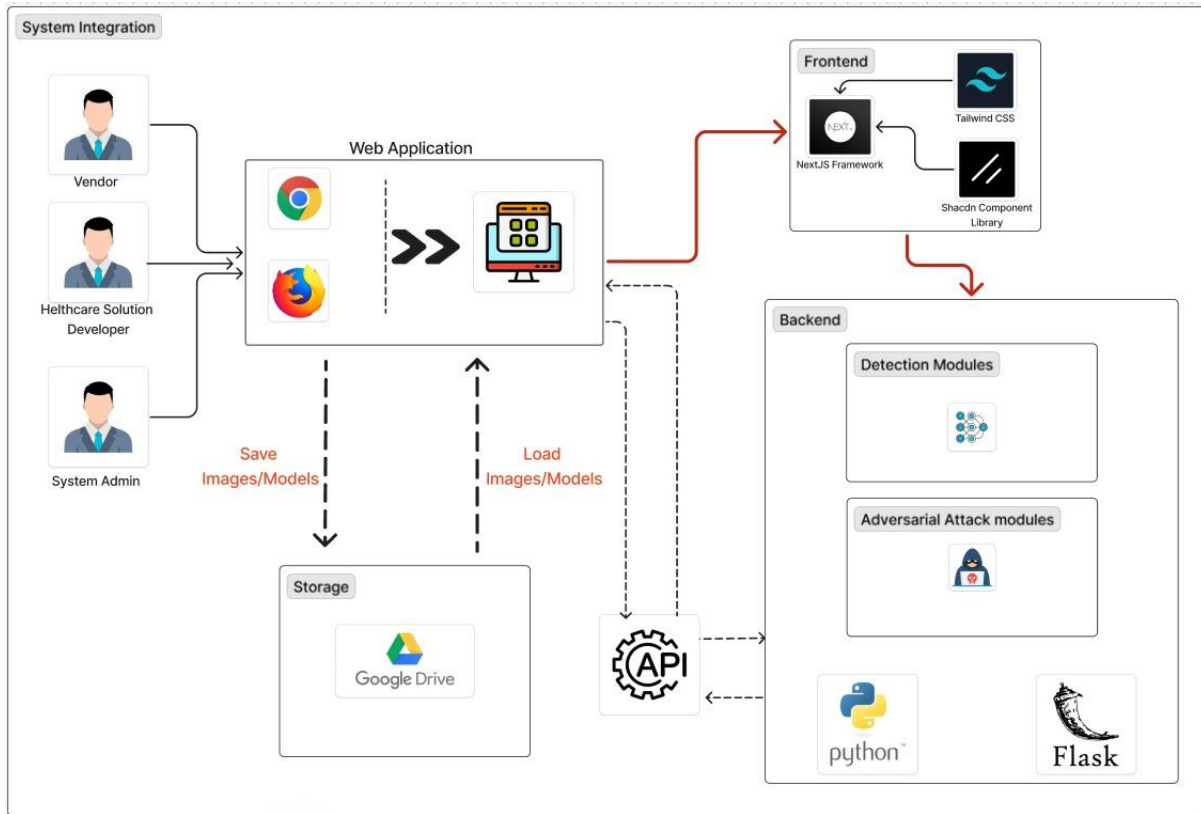


Figure 11: System Integration

3.1 System Overview

Protecting sensitive healthcare data from increasingly complex hostile assaults is the goal of the developed system, which is a complete and elaborately constructed cybersecurity platform specifically made for healthcare applications. To provide strong cybersecurity across a range of operational domains in healthcare settings, this system is made up of several integrated components, each of which is intended to perform unique but related functions.

1. Frontend Infrastructure

The system's frontend is crafted to maximize user engagement and operational effectiveness. The frontend provides responsive, scalable, and high-performance web application experience through the use of NextJS. NextJS speeds up the application's rendering, makes server-side rendering easier, and greatly enhances the user experience in general. In addition to NextJS, the frontend uses Tailwind CSS, which is well-known for its comprehensive, utility-first approach to CSS and allows for quick UI creation and uniformity throughout the application. Additionally, the system's interface is improved by the inclusion of the Shadcn Component Library, which offers advanced UI components that support user-friendliness and straightforward navigation.

2. Backend Infrastructure

Flask, a very flexible Python-based micro web framework known for its lightweight design and broad customization options, powers the integrated system's backend. Real-time image processing, data analysis, model execution, and threat detection are just a few of the intricate, computationally demanding activities that Flask makes it easier to handle. Its adaptable design facilitates the smooth integration of several machine learning models, each of which is set up to carry out detection and analysis tasks, improving the overall effectiveness of cybersecurity.

3. Detection Modules

The system's advanced detection modules are the foundation of its cybersecurity capabilities. These modules use sophisticated neural network architectures to conduct thorough, in-depth evaluations of images relevant to healthcare in order to spot tiny hostile modifications that could jeopardize patient safety or diagnostic accuracy. Incoming data is continuously analyzed by detection modules, which quickly spot possible dangers and notify stakeholders in a timely manner. By enabling prompt and proactive mitigation techniques, these warnings greatly lower the danger posed by hostile cyberthreats.

4. Adversarial Attack Modules

The integrated system includes adversarial attack modules to thoroughly verify the detection modules' resilience. These modules extensively stress-test the system's defenses and offer vital insights into potential weaknesses by simulating complex adversarial attacks in controlled situations. By taking a proactive stance, the system's cybersecurity defenses are constantly assessed and strengthened, preserving its resistance to new online dangers.

5. API and Communication Interfaces

In order to provide smooth integration and communication between frontend interfaces, backend computing modules, and external data storage solutions, the system makes use of a strong API infrastructure that acts as the essential communication backbone. The API is thoughtfully designed to manage intricate data transactions, such as the safe transfer, archiving, and retrieval of machine learning models and data. Furthermore, the API infrastructure guarantees data encryption and secure authentication, preserving the availability, confidentiality, and integrity of critical medical data.

6. Cloud Storage and Data Management

The system uses Google Drive as its main cloud storage option for safe, scalable, and effective data storage. The strong infrastructure of Google Drive offers high availability, large storage space, safe data encryption, and efficient data retrieval procedures. Google Drive greatly improves the operational efficiency and cybersecurity of the system by safely storing and maintaining crucial healthcare imaging and machine learning models.

7. Workflow Coordination

A carefully planned workflow that balances the interactions between frontend user interfaces, backend computational operations, API communications, and cloud storage solutions is at the heart of the integration architecture. Effective operational flow, quick threat detection, strong cybersecurity defenses, and a markedly enhanced user experience are all guaranteed by this well-coordinated integration strategy. Throughout the integrated cybersecurity environment, the coordination strategy fosters smooth interoperability, real-time responsiveness, and optimal operational performance.

In summary, the intricate integration of these elements results in a complex, all-encompassing cybersecurity platform that is specially equipped to handle the requirements and difficulties that healthcare systems face today. By using a comprehensive approach, the system greatly improves cybersecurity protections, guaranteeing defense against new threats and maintaining the dependability and operational integrity of healthcare applications.

4. Commercialization

Our commercialization strategy converts the results of our adversarial robustness research into a comprehensive, vendor-ready validation platform specifically designed for medical imaging AI. The platform fundamentally provides a modular, web-based environment that enables users, including AI developers, medical software suppliers, and hospital IT departments, to upload their CNN models and subject them to a series of hostile tests. These encompass style-transfer manipulation, structure-invariant perturbations, gradient-norm attacks, and mask-based distortions. Through the automation of adversarial example production, the platform provides comprehensive performance reports—encompassing accuracy, precision, recall, F1-scores, confusion matrices, and security scoring dashboards—enabling stakeholders to identify precisely where and how their diagnostic models falter.

To facilitate uptake, we will market this platform as a Software-as-a-Service (SaaS) subscription, featuring tiered options to accommodate various organizational requirements. Fundamental plans will encompass standard adversarial test suites and performance metrics, whereas premium tiers will incorporate real-time API integration for CI/CD pipelines, support for additional imaging modalities (CT, MRI, ultrasound), and customizable security scoring in accordance with regulatory frameworks (e.g., FDA, CE marking). An enterprise version will encompass on-premise deployment, dedicated SLAs, and professional services—including customized test scenarios, model hardening workshops, and compliance consulting—facilitating large healthcare systems in the seamless integration of adversarial validation into their AI governance frameworks.

Our future plans encompass strategic alliances with established medical-device integrators, cloud infrastructure providers, and regulatory consultants to guarantee scalability, interoperability, and compliance. Proposed feature enhancements—including federated adversarial benchmarking among several institutions, sophisticated visualization tools for threat modeling, and automated remedial recommendations—will further distinguish our product. Our commercialization strategy integrates advanced AI security research with practical tools and services, addressing a significant unmet need in healthcare while creating a sustainable, high-value enterprise that promotes the secure implementation of deep learning in clinical settings.

4.1 Stake Holders

The integrated cybersecurity system has been carefully crafted to meet the needs and demands of multiple important stakeholders, each of whom has specific duties and responsibilities that are essential to the overall cybersecurity framework's success. To guarantee seamless system deployment, operational effectiveness, and strong cybersecurity measures, these stakeholders must be effectively identified, engaged, and managed. System administrators, vendors, and healthcare solution developers have been highlighted as the main players for this cybersecurity integration.

1. Vendors

One of the main stakeholders in the integrated system is vendors. Their primary responsibility is to supply vital resources including medical imaging, updated machine learning models, and critical cybersecurity updates. Vendors have a major impact on the operational reliability of the system by guaranteeing the resources' constant availability, correctness, and compatibility. Among their duties is the methodical provisioning of resources through the safe uploading of machine learning models and images pertaining to healthcare through a specific web interface. Vendors assure smooth integration and effective operation by making sure these resources continuously satisfy the system's strict standards and compatibility requirements. Additionally, vendors are essential in providing timely updates and continuous technical support for cybersecurity components, including the release of crucial security patches and seamless resource integration. Vendors must closely follow cybersecurity guidelines to guarantee data security, confidentiality, and integrity. They want transparent operational standards, safe data upload procedures, easy-to-use resource management interfaces, and well-defined support and troubleshooting contact routes.

2. Healthcare Solution Developers

Another essential stakeholder group for incorporating cybersecurity modules into healthcare systems is healthcare solution developers. One of their main responsibilities is to seamlessly integrate adversarial cybersecurity and sophisticated detection modules into already-existing healthcare apps. Developers must ensure that integration does not adversely affect the clinical accuracy of the healthcare applications, user experience, or overall system performance. Additionally, they must balance advanced cybersecurity protections with the pragmatic realities of clinical operations by tailoring cybersecurity solutions to meet application-specific requirements. To guarantee that the systems react to attacks promptly and efficiently without creating latency or operational limits, healthcare solution developers actively optimize the performance of cybersecurity modules. To guarantee that cybersecurity modules function precisely and dependably in actual healthcare settings, they are also in charge of thorough testing and validation. In addition to clear instructions on how to comply with healthcare cybersecurity compliance standards, developers anticipate thorough and accurate documentation to aid integration processes, effective technical support mechanisms, low system overhead from cybersecurity modules, and strong interoperability with current healthcare infrastructures.

3. System Administrators

A crucial stakeholder group, system administrators are in charge of monitoring the overall security, performance, health, and compliance with regulations of the integrated cybersecurity solution. Their responsibilities include thorough system monitoring to guarantee real-time visibility into user activity, system health data, and cybersecurity events. Administrators keep situational awareness and proactively address possible dangers by utilizing the system's sophisticated analytical capabilities and dedicated dashboards. In order to prevent unauthorized use of vital resources, system administrators additionally oversee access control, making sure that only authorized individuals can access key cybersecurity modules, data, and application capabilities. They oversee

responding to cybersecurity events quickly and efficiently by implementing pre-established procedures and closely collaborating with other stakeholders. Additionally, by routinely assessing system operations and putting the required enhancements or corrective actions into place, system administrators actively guarantee adherence to regulatory standards unique to the healthcare industry. They are also responsible for doing routine maintenance tasks that preserve system availability, dependability, and security integrity, such as software updates, security patches, system backups, and performance optimization. Strong support infrastructure, clear and actionable incident management procedures, extensive documentation for cybersecurity best practices and regulatory compliance, sophisticated real-time analytics and reporting features, and strong yet user-friendly management tools are all things that administrators look for.

5. Conclusion

This study offers a comprehensive evaluation of the susceptibility of CNN-based medical imaging models to advanced adversarial attacks, with a particular emphasis on the PGN technique. The results confirm that PGN-generated perturbations, by focusing on flat regions of the loss surface, significantly improve the transferability of adversarial examples, making them harder to defend against. Among the CNN models tested, each exhibited varying degrees of vulnerability, with deeper architectures like ResNet50 showing relatively stronger resilience. However, the findings underscore that no model is inherently immune to such attacks.

Defense mechanisms, notably adversarial training, proved effective in restoring classification accuracy while maintaining the clinical relevance of images. Other defenses, such as JPEG compression and high-level denoising, provided partial protection but often compromised image fidelity. These insights are particularly crucial for medical AI developers, as they demonstrate the importance of integrating robust, context-aware defenses during model development rather than post-deployment.

Overall, this research highlights a pressing need to continuously evaluate and fortify deep learning systems against evolving adversarial threats. By simulating and defending against PGN and related attacks within a clinically relevant framework, this work contributes to the advancement of secure and reliable ML solutions in healthcare, promoting trust, accuracy, and safety in automated diagnostic processes.

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] Z. Z. J. Z. Xiaosen Wang, Structure Invariant Transformation for better Adversarial Transferability, 2024: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France.
- [4] H. L. X. W. F. S. Y. L. Zhijin Ge, Boosting Adversarial Transferability by Achieving Flat Local Maxima, Vancouver, Canada: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.
- [5] Z. Y. Q. L. z. L. Tao Wang, Boost Adversarial Transferability by Uniform Scale and Mix Mask Method, 2023.
- [6] 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.
- [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] 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.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] 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.

- [13] T. S. a. P. S. W. a. A. R. K. a. K. V. P. a. K. P. G. a. C. P. Arulananth, Classification of Paediatric Pneumonia Using Modified DenseNet-121 Deep-Learning Model, IEEEAccess, 2024.
- [14] C. Singh and A. Singh, Deep Learning-Based Architectures for RLDD-Rice Leaf Disease Detection, Dehradun, India: 2024 International Conference on Cybernation and Computation (CYBERCOM), 2024.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] E. Jain and S. Choudhary, Enhancing Tuberculosis Diagnosis with DenseNet121 and Grad-CAM: A Deep Learning Approach for Accurate and Interpretable Chest X-ray Analysis, Manama, Bahrain: 2024 International Conference on Decision Aid Sciences and Applications (DASA), 2025.
- [19] 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.
- [20] A. T. 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.
- [21] 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.
- [22] S. Tammina, Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Image, California, San Diego: International Journal of Scientific and Research Publications, 2019.
- [23] Y. G. Jiahui Tao, J. Sun, Y. Bie and H. Wang, Research on vgg16 convolutional neural network feature classification algorithm based on Transfer Learning, Shanghai, China: 2021 2nd China International SAR Symposium (CISS), 2021.
- [24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, Rethinking the Inception Architecture for Computer Vision, Las Vegas, NV, USA: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.