

Assessing CNN Robustness in Medical Imaging Systems: Adversarial Threats and Defensive Measures

Janith Sandamal

IT21166860

BSc (Hons) in Information Technology Specializing in Cyber Security

Department of Computer System Engineering

Sri Lankan Institute of Information and Technology

April 2025

Assessing CNN Robustness in Medical Imaging Systems: Adversarial Threats and Defensive Measures

P.G.E Janith Sandamal

IT21166860

BSc (Hons) in Information Technology Specializing in Cyber Security

Department of Computer System Engineering

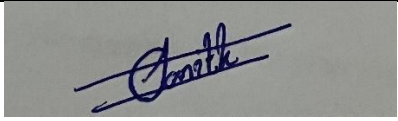
Sri Lankan Institute of Information and Technology

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 articles or books).

Name	Student ID	Signature
Sandamal P.G.E. J	IT21166860	

The above candidate has carried out research for the bachelor's degree Dissertation under my supervision.

Signature of the Supervisor:

Date:

ABSTRACT

In the modern healthcare ecosystem, deep learning models particularly Convolutional Neural Networks (CNNs) are widely deployed to assist clinical decision-making through medical image analysis. These models have demonstrated high diagnostic accuracy in detecting diseases such as pneumonia and pneumothorax from chest X-ray images. However, a critical yet often overlooked vulnerability persists: their susceptibility to adversarial attacks. These attacks involve subtle, carefully crafted perturbations that can deceive neural networks into making incorrect predictions, potentially leading to misdiagnoses. Alarming, many healthcare system vendors and AI solution developers remain unaware of the growing threat landscape surrounding adversarial machine learning.

This research investigates the Structure Invariant Attack (SIA) an emerging adversarial strategy that applies block-wise transformations to chest X-rays without visibly altering the clinical appearance of the image. Focusing on the DenseNet121 architecture, we evaluate how SIA manipulates model behavior and compromises diagnostic performance across three medical conditions: Normal, Pneumonia, and Pneumothorax.

To counter this threat, we propose and rigorously test two lightweight defense mechanisms: JPEG compression and Total Variation Minimization (TVM). In addition, we develop a preliminary detection method capable of identifying adversarial manipulations either prior to or post-classification. Our findings show that the SIA attack notably degrades classification accuracy, particularly by targeting high-confidence predictions. However, integrating defense mechanisms significantly restores model performance and mitigates the attack's impact.

This study provides a high-level yet actionable overview of adversarial vulnerabilities, tailored for healthcare AI stakeholders. By illustrating both the attack vector and corresponding defensive strategies, this work offers a foundational framework for enhancing the security, reliability, and robustness of CNN-based medical imaging systems. It aims to inform and empower system integrators, developers, and researchers in the healthcare domain to proactively address adversarial threats in real-world clinical environments.

Keywords: CNN Robustness, Adversarial Attack, DenseNet121, Structure Invariant Attack

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to everyone who contributed to the successful completion of this research project.

First and foremost, I extend my deepest appreciation to my supervisor, Dr. Harinda Fernando, and Co-Supervisor Mr. Kavinga Yapa, for their invaluable guidance, constructive feedback, and continuous support throughout the research journey. Their expert insights played a crucial role in shaping the direction of this study and ensuring its academic integrity.

I am also thankful to the academic and technical staff of the Computer Systems and Engineering Department for providing access to resources, infrastructure, and mentoring that enabled smooth execution of each research phase. Their commitment to academic excellence inspired and motivated me to pursue this topic rigorously.

Special thanks go to my teammates and peers in the research group for their collaboration, knowledge sharing, and encouragement during complex implementation stages, particularly when testing and integrating adversarial attack and defense components.

I would also like to acknowledge the developers and maintainers of open-source tools and frameworks such as TensorFlow, Keras, NumPy, OpenCV, Portaniner and Google Colab, without which the technical foundation of this research would not have been possible.

Finally, I am grateful to my family and friends for their patience, moral support, and constant encouragement, which kept me focused and motivated throughout this challenging but fulfilling research experience.

TABLE OF CONTENTS

DECLARATION	2
ABSTRACT	3
ACKNOWLEDGEMENT.....	4
TABLE OF CONTENTS	5
LIST OF FIGURES	7
LIST OF ABBREVIATIONS.....	8
INTRODUCTION	10
1.1. Role of Deep Learning in Medical Imaging.....	10
1.2. How Deep Learning Works In Medical Imaging	10
1.3. Applications in Real-World Medical Systems	11
1.3.1. Advantages in healthcare.....	11
1.4. Importance of Security in AI-Powered Healthcare Systems	12
1.5. Adversarial Threats in Medical Imaging	13
1.5.1. How adversarial attacks work	13
1.5.2. Why medical imaging is a prime target.....	14
1.5.3. Need for awareness and countermeasures	14
1.6. Structure Invariant Attack (SIA): A Novel Threat Vector	14
1.7. Overview of DenseNet121 and Its Use in Healthcare.....	15
1.7.1. Skip-connection advantage	15
1.7.2. Use in healthcare applications	16
1.7.3. Why densenet121 was chosen for this research	16
1.7.4. Strengths and limitations under adversarial threats	16
1.8. Background and Literature Survey.....	18
1.8.1. Cnn performance in medical image classification.....	18
1.8.2. Studies on adversarial attacks in healthcare	18
1.8.3. Existing defense strategies.....	18
1.9. Research Gap	19
1.10. Research Problem.....	21
2. OBJECTIVES	22
2.1. Core Research Objectives.....	22
3. METHODOLOGY	23
3.1. Research and Technical Requirements.....	23
3.1.1. Research scope and objectives mapping.....	23

3.1.2.	Dataset requirements	24
3.1.3.	Model and environment requirements	25
3.2.	Functional Requirements.....	25
3.3.	Non-Functional Requirements.....	26
3.4.	Budget Requirements	27
3.5.	Feasibility Study	27
3.5.1.	Technical feasibility	27
3.6.	Design.....	29
3.6.1.	Design philosophy.....	29
3.6.2.	System workflow	29
3.6.3.	Model-level design choices.....	31
3.7.	System Implementation.....	31
3.7.1.	Model and attack implementation	31
3.7.2.	Defense and detection strategy	32
3.8.	Technology Stack.....	34
3.9.	Results & Discussion.....	35
3.9.1.	Baseline model evaluation (pre-attack)	35
3.9.2.	Adversarial impact analysis (post-sia attack)	37
3.9.3.	Defense evaluation (post jpeg + tvn)	38
3.9.4.	Detection system evaluation	40
4.	COMMERCIALIZATION	42
5.	DESCRIPTION OF PERSONAL COMPONENT.	43
6.	CONCLUSION	44
7.	REFERENCES.....	45
8.	APPENDICES	46
	Appendix A - Data Preprocessing and Dataset prepare	46
	Appendix B - Fine-tune DenseNet121 Architecture.....	48
	Appendix C - Structure Invariant Attack Algorithm.....	49
	Appendix D - SIA Attack hybrid defense measure algorithm.....	50
	Appendix E - SIA attack detection algorithm	50

LIST OF TABLES

Table 1: Pros and Cons of DenseNet121 in Medical Imaging Under Adversarial Conditions.....	17
Table 2: Mapping Research Objectives to System Modules	23
Table 3: Functional Requirements and Modules	26
Table 4: Technical Feasibility Assessment Summary	28

LIST OF FIGURES

Figure 1: Visual Representation of CNN Feature Extraction in Chest X-rays.....	10
Figure 2: AI-Augmented Chest X-ray Diagnosis Workflow	11
Figure 3: Risk Landscape of AI in Healthcare	12
Figure 4: Visual Comparison of Original vs. Adversarial Image under SIA Attack	13
Figure 5:DenseNet121 Architecture Overview	16
Figure 6: Research Gap.....	20
Figure 7: Adversarial Threat impact in Medical Imaging.....	21
Figure 8: NIH dataset classes and image count per class	24
Figure 9: Balanced NIH Dataset.....	24
Figure 10: Portainer Infrastructure	25
Figure 11: Overall System Design	30
Figure 12: Individual Component Design	30
Figure 13: Confusion Matrix (Baseline Model).....	35
Figure 14: Model Training and Validation Accuracy/Loss	36
Figure 15: Class-wise Accuracy (Baseline)	36
Figure 16: Visualization: Original vs. SIA-Attacked Image	37
Figure 17: Confusion Matrix After SIA Attack	37
Figure 18: Accuracy Drop per Class (Post-Attack)	38
Figure 19: Visual Comparison: Original vs. SIA vs. Defended Image.....	38
Figure 20: Confusion Matrix After Defense	39
Figure 21: Class-wise Accuracy (Post-Defense).....	39
Figure 22: Detection Accuracy Chart	40
Figure 23: Detection Demonstration Sample	40
Figure 24: Detection Success Rate (Visual Summary)	41
Figure 25: System Integration Design	42

LIST OF ABBREVIATIONS

Abbreviation	Description
AI	Artificial Intelligence
API	Application Programming Interface
CNN	Convolutional Neural Network
DL	Deep Learning
ML	Machine Learning
SIA	Structure Invariant Attack
TVM	Total Variation Minimization
JPEG	Joint Photographic Experts Group (Image Compression Standard)
FGSM	Fast Gradient Sign Method (Adversarial Attack)
PGD	Projected Gradient Descent (Adversarial Attack)
C&W	Carlini & Wagner (Adversarial Attack)
XAI	Explainable Artificial Intelligence
GPU	Graphics Processing Unit
NIH	National Institutes of Health
PACS	Picture Archiving and Communication System
TV	Total Variation
IoU	Intersection over Union
AUC	Area Under the Curve
ROC	Receiver Operating Characteristic
SGD	Stochastic Gradient Descent
ReLU	Rectified Linear Unit
L2	L2 Regularization (Weight Decay)
FP	False Positive
FN	False Negative

TP	True Positive
TN	True Negative
X-ray	Radiographic Imaging using X-rays
Colab	Google Colaboratory
OpenCV	Open Source Computer Vision Library
Keras	High-level Neural Network API (runs on top of TensorFlow)
TensorFlow	Open-source Machine Learning Framework by Google
SciPy	Scientific Python Library
NumPy	Numerical Python Library
ROC Curve	Receiver Operating Characteristic Curve

INTRODUCTION

1.1. Role of Deep Learning in Medical Imaging

Deep learning has emerged as a transformative technology in the field of medical imaging, offering unprecedented capabilities for disease detection, classification, and diagnosis. At the heart of this revolution are Convolutional Neural Networks (CNNs), which have demonstrated exceptional performance in analyzing complex visual data such as chest X-rays, CT scans, and MRIs. These models can learn intricate patterns in pixel-level data and translate them into clinically meaningful predictions, often at a level comparable to experienced radiologists.

Traditional diagnostic workflows rely heavily on the expertise of medical professionals, who manually review imaging scans to identify signs of abnormalities. However, human analysis can be time-consuming, resource-intensive, and susceptible to fatigue or diagnostic variation. Deep learning models address these challenges by offering rapid, consistent, and automated interpretation of medical images, thereby augmenting clinical decision-making.

1.2. How Deep Learning Works In Medical Imaging

CNNs are specifically designed to work with image data. They consist of layers that automatically detect visual features, such as edges, textures, and shapes, which become increasingly abstract as the network goes deeper. For instance, in the context of chest X-rays, early layers may identify basic structures like rib outlines or lung boundaries, while deeper layers recognize high-level patterns indicative of diseases like pneumonia or pneumothorax.

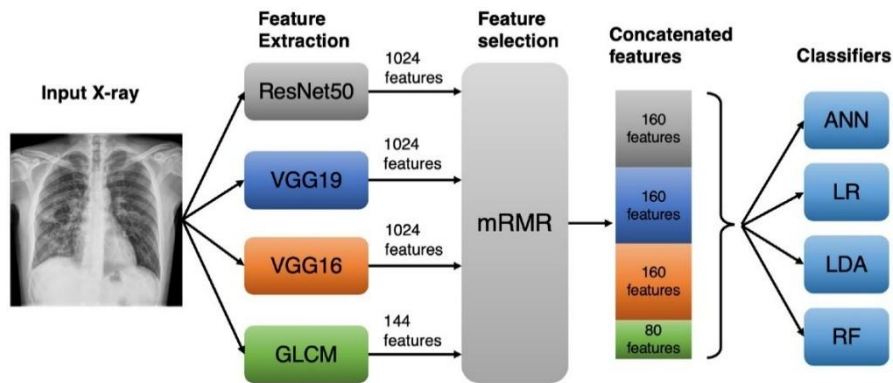


Figure 1: Visual Representation of CNN Feature Extraction in Chest X-rays

This layered abstraction allows deep learning models to identify subtle patterns that may be missed by the human eye, making them invaluable in early disease detection, especially in resource-limited settings where radiologists are scarce.

1.3. Applications in Real-World Medical Systems

The integration of deep learning into medical imaging is no longer experimental—it is actively used in clinical environments. Several FDA-approved tools already assist radiologists in detecting lung nodules, fractures, brain hemorrhages, and breast cancer. Chest X-rays, being one of the most commonly performed radiological procedures globally, are a primary focus for deep learning applications due to their diagnostic relevance and standardized format.

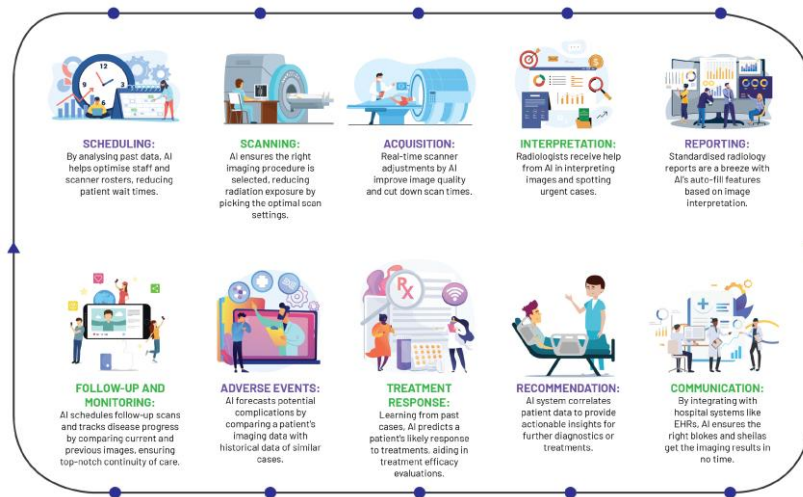


Figure 2: AI-Augmented Chest X-ray Diagnosis Workflow

1.3.1. Advantages in healthcare

- **Speed and Efficiency:** AI can analyze thousands of images within minutes, accelerating diagnosis and patient triage.
- **Scalability:** Deep learning systems can be deployed across multiple clinics, helping to bridge the gap in areas with limited medical expertise.
- **Consistency:** Unlike human interpretation, AI models provide consistent results, reducing the risk of diagnostic variability.
- **Early Detection:** CNNs can detect minute anomalies before they become clinically evident, enabling proactive intervention.

However, the adoption of deep learning is not without its concerns. While the performance of these models continues to improve, questions around their interpretability, data bias, and security remain critical. One of the most overlooked but significant threats is the model's vulnerability to adversarial attacks subtle manipulations to input images that can cause AI to misclassify, often without any visible change to human observers.

This challenge forms the core focus of the present research, which explores how deep learning models, particularly DenseNet121, behave under adversarial conditions such as the Structure Invariant Attack (SIA), and how such vulnerabilities can be mitigated to ensure the safe and trustworthy deployment of AI in medical imaging.

1.4. Importance of Security in AI-Powered Healthcare Systems

As artificial intelligence (AI) continues to integrate into modern healthcare systems, its role in diagnostics, patient management, and clinical decision support has become indispensable. However, this integration introduces new security challenges that, if unaddressed, could jeopardize patient safety, data integrity, and institutional trust.

AI models, particularly those used in medical imaging, operate in critical environments where any misdiagnosis can have life-altering consequences. While these systems are designed to enhance efficiency and accuracy, they also present an expanded attack surface for cyber threats. Among these, adversarial attacks where input images are subtly manipulated to deceive AI models—pose a serious risk. Such perturbations can lead to incorrect diagnoses, even though the image appears unchanged to human observers.

Healthcare data is also highly sensitive and protected under regulations such as HIPAA and GDPR. AI systems that relieve this data must be resilient against both external attacks and internal misuse. Failure to secure these models could result in data breaches, unauthorized model manipulation, and loss of clinical credibility.

Moreover, the black-box nature of many deep learning models makes it difficult to detect when they are under attack or behaving abnormally. This lack of transparency can delay response times and increase the risk of undetected failures in diagnostic workflows.

Therefore, ensuring the security, robustness, and interpretability of AI systems is not optional, it is essential. Building secure AI requires a multidisciplinary approach that combines adversarial awareness, defensive modeling techniques, and proactive monitoring mechanisms. This is especially vital for developers and system integrators who design and deploy medical AI in real-world clinical settings.

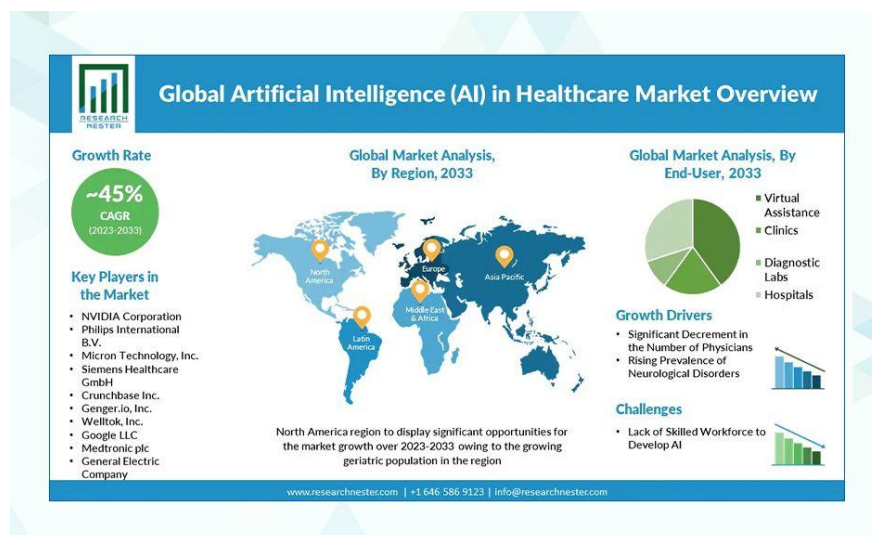


Figure 3: Risk Landscape of AI in Healthcare

1.5. Adversarial Threats in Medical Imaging

Adversarial machine learning is an emerging field that exposes a critical vulnerability in deep learning systems: the ability to be deceived by imperceptible input manipulations. These specially crafted perturbations, known as adversarial examples, can cause a well-trained Convolutional Neural Network (CNN) to make incorrect predictions—without any obvious visual change to the input image from a human perspective.

In the context of medical imaging, where CNNs are increasingly used to detect and classify diseases from radiographs, MRIs, or CT scans, the implications of adversarial attacks are particularly alarming. A single modified pixel pattern or structural distortion in a chest X-ray, for instance, can lead the model to misclassify a pneumothorax as a normal lung, or misdiagnose pneumonia ultimately risking incorrect treatment or delayed medical intervention.

1.5.1. How adversarial attacks work

These attacks exploit the mathematical structure of neural networks. By introducing carefully calculated noise into the image often at a magnitude so small that it is invisible to the human eye the attacker shifts the model's decision boundary just enough to produce an incorrect output with high confidence.

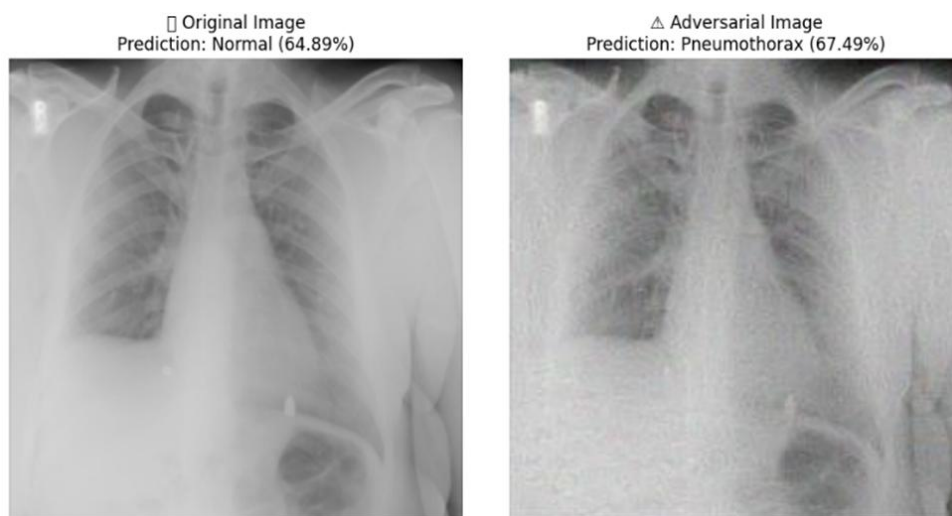


Figure 4: Visual Comparison of Original vs. Adversarial Image under SIA Attack

Unlike typical software vulnerabilities, adversarial attacks are model-agnostic they do not require access to the underlying code or training data. This makes it difficult to detect and prevent using traditional security mechanisms. In black-box settings, attackers can even transfer adversarial perturbations across models, amplifying the threat in real-world deployments.

Real-world implications

Several academic studies have demonstrated that deep learning models used in dermatology, ophthalmology, and radiology can be consistently fooled with adversarial inputs. For instance, a CNN trained to detect lung

nodules may falsely report “no findings” when adversarial noise is added despite the actual presence of a mass in the image.

In real-world deployments:

- Attackers could potentially manipulate images during transmission or storage (e.g., in PACS or cloud-based diagnostic platforms).
- Malicious insiders or compromised devices could subtly alter image data before it reaches diagnostic tools.
- Autonomous diagnostic systems may unknowingly act on incorrect predictions, affecting clinical workflows.

1.5.2. Why medical imaging is a prime target

- High reliance on automated interpretation
- Lack of robustness validation in model deployment pipelines
- Minimal human oversight in triage systems
- Assumption of image integrity in standard radiology workflows

These characteristics make adversarial threats not just a technical curiosity, but a tangible concern for clinical AI safety.

1.5.3. Need for awareness and countermeasures

Unfortunately, many medical AI developers and system integrators are not yet aware of these risks. As the use of AI becomes ubiquitous in healthcare diagnostics, it is crucial to understand and address adversarial threats during both model development and system integration stages.

This research takes a step forward by simulating a novel adversarial threat Structure Invariant Attack (SIA) and demonstrating its impact on a state-of-the-art medical CNN (DenseNet121). Additionally, it proposes defenses and detection strategies to improve the model’s resilience, aiming to inform and empower healthcare developers toward building safer AI systems.

1.6. Structure Invariant Attack (SIA): A Novel Threat Vector

The Structure Invariant Attack (SIA) represents a new class of adversarial threats specifically designed to deceive deep learning models by exploiting their reliance on structural patterns within images. Unlike traditional pixel-level adversarial attacks which introduce minute noise uniformly across the image SIA operates by applying block-wise transformations that alter the internal structure of the image while preserving its overall visual appearance.

In SIA, the image is divided into smaller blocks, and controlled manipulations such as shuffling, rotation, or localized distortions are applied within or between these blocks. These changes are subtle enough that a human observer, including trained medical professionals, would likely perceive the image as unchanged or diagnostically acceptable. However, for CNN-based models, which are highly sensitive to spatial arrangements and feature hierarchies, such structural alterations can severely degrade classification accuracy or lead to confidently incorrect predictions.

What makes SIA particularly dangerous in medical imaging is its stealthiness and high transferability. Standard input validation techniques or adversarial detectors—often trained to spot noisy perturbations fail to flag SIA-modified inputs because the overall pixel distribution and image texture remain statistically consistent with benign data.

In diagnostic scenarios where clinicians depend on AI outputs to support or prioritize decision-making, an undetected SIA attack could silently cause incorrect triage or diagnosis, putting patient safety at risk.

This research specifically investigates the vulnerability of DenseNet121 to SIA in chest X-ray classification and evaluates countermeasures to enhance the model's robustness against such stealthy, structure-oriented adversarial manipulations.

1.7. Overview of DenseNet121 and Its Use in Healthcare

DenseNet121 (Densely Connected Convolutional Network) is a state-of-the-art deep learning architecture designed to enhance feature reuse, reduce model complexity, and improve gradient flow during training. It is particularly known for its dense connectivity pattern, where each layer receives inputs from all preceding layers and passes its output to all subsequent layers. This design leads to stronger feature propagation, efficient parameter usage, and improved performance, especially on medical image datasets where subtle features matter.

1.7.1. Skip-connection advantage

Unlike traditional CNNs where information passes sequentially layer by layer, DenseNet121 employs skip connections that allow layers to directly access outputs from earlier layers. This has three critical benefits:

- Better gradient flow during backpropagation, reducing the vanishing gradient problem.
- Efficient feature reuse, enabling the model to learn both low-level and high-level representations more effectively.
- Parameter efficiency, as it uses fewer parameters than other deep models of similar depth.

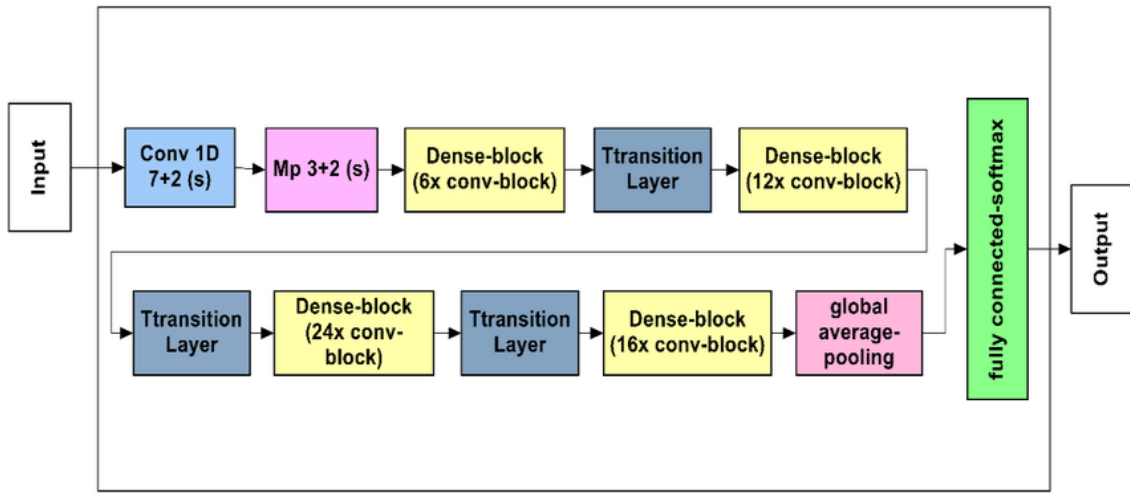


Figure 5: DenseNet121 Architecture Overview

1.7.2. Use in healthcare applications

DenseNet121 has demonstrated outstanding performance in medical imaging tasks, including:

- Chest X-ray classification (e.g., detecting pneumonia, tuberculosis, pneumothorax)
- Retinal disease detection in fundus photography
- Tumor classification in histopathological images

Its ability to extract fine-grained features makes it especially well-suited for radiological tasks where minute abnormalities such as lesions, opacities, or nodules may be diagnostically significant.

1.7.3. Why densenet121 was chosen for this research

- Proven performance in prior chest X-ray studies (e.g., NIH ChestX-ray14, CheXpert)
- High resolution feature preservation, crucial for detecting small anomalies in grayscale medical images
- Adaptability for transfer learning using pre-trained ImageNet weights, saving training time

1.7.4. Strengths and limitations under adversarial threats

Strengths:

- Strong generalization when trained with diverse and augmented data
- Stability across varying image quality and resolutions

Limitations:

- Vulnerable to adversarial attacks like SIA, which exploit the very feature hierarchies DenseNet relies on
- Dense connections, while useful for learning features, can propagate adversarial perturbations more effectively throughout the network

Table 1: Pros and Cons of DenseNet121 in Medical Imaging Under Adversarial Conditions

Feature	Advantage	Limitation
Dense connectivity	Enhanced feature reuse	Potential amplification of adversarial patterns
Parameter efficiency	Lower overfitting risk	Reduced capacity for defense layering
Pretrained model availability	Easier fine-tuning on medical datasets	May transfer adversarial vulnerabilities
High-resolution processing	Suitable for subtle feature detection	Sensitive to structural input modifications

1.8. Background and Literature Survey

1.8.1. Cnn performance in medical image classification

Convolutional Neural Networks (CNNs) have become pivotal in medical image analysis, demonstrating exceptional performance in tasks such as disease detection, segmentation, and classification. Their hierarchical architecture enables the extraction of complex features from medical images, facilitating accurate diagnoses. For instance, Kermany et al. achieved 92% accuracy in pneumonia detection using transfer learning on a small X-ray dataset. Similarly, CNNs have been effectively applied to detect tumors, skin lesions, and other anomalies[2].

The adoption of transfer learning has further enhanced CNN performance, especially when dealing with limited medical datasets. By leveraging pre-trained models on large datasets like ImageNet, researchers can fine-tune CNNs for specific medical imaging tasks, reducing the need for extensive labeled data [4].

1.8.2. Studies on adversarial attacks in healthcare

Despite their success, CNNs are vulnerable to adversarial attacks subtle, often imperceptible perturbations to input images that can lead to incorrect predictions. In the healthcare domain, such attacks pose significant risks, potentially leading to misdiagnoses. Research has shown that medical image analysis systems can be deceived by these perturbations, undermining their reliability [1].

Adversarial attacks in healthcare are particularly concerning due to the high stakes involved. Misclassifications resulting from such attacks can lead to inappropriate treatment or missed diagnoses, emphasizing the need for robust defense mechanisms [6].

1.8.3. Existing defense strategies

To counter adversarial attacks, several defense strategies have been proposed:

- **JPEG Compression:** This technique reduces image quality to eliminate adversarial noise. While effective against certain attacks, it may also degrade important diagnostic features [3].
- **Total Variation Minimization (TVM):** TVM aims to smooth images, removing perturbations while preserving essential structures. However, excessive smoothing can obscure critical details [15].
- **Adversarial Training:** Incorporating adversarial examples during training enhances model robustness. Although effective, this approach increases computational demands and may not generalize well to unforeseen attacks [7].

Each method has its advantages and limitations, and often, a combination of strategies is employed to achieve optimal protection.

1.9. Research Gap

Despite significant advancements in the application of deep learning for medical imaging, the robustness and security of Convolutional Neural Networks (CNNs) against adversarial threats remain underexplored particularly in clinical settings. While adversarial attacks such as Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Carlini & Wagner (C&W) have been extensively studied in the computer vision domain, their specific implications in medical diagnostics have only recently garnered attention.

However, one emerging adversarial approach the Structure Invariant Attack (SIA) has received minimal investigation, especially in the context of healthcare. SIA introduces block-wise spatial transformations to the input image, subtly altering its internal structure while preserving its visual appearance to human observers. This attack poses a unique threat in medical imaging, where CNNs rely heavily on the spatial arrangement of anatomical features for disease classification. Despite its potential for real-world exploitation, SIA has not yet been evaluated on medical datasets in any existing literature to date.

Furthermore, there is a notable lack of lightweight and real-time defensive and detection strategies tailored for such structure-preserving adversarial attacks. Existing defenses like adversarial training and Total Variation Minimization (TVM) are either computationally intensive or not optimized for practical deployment in clinical pipelines. Detection methods, where available, are often tailored to noise-based perturbations and are ineffective against block-wise manipulations that preserve global pixel statistics.

Another critical oversight in current research is the absence of a systematic evaluation of the DenseNet121 architecture under SIA conditions. Although DenseNet121 is widely adopted for chest X-ray classification due to its high accuracy and efficient feature reuse, its vulnerability to structure-based adversarial attacks has not been explored, nor have its defensive capabilities been benchmarked in such scenarios.

This research directly addresses these gaps by:

- Implementing the SIA attack on real-world chest X-ray images for the first time,
- Evaluating its impact on DenseNet121's classification performance,
- Proposing and testing lightweight defense mechanisms such as JPEG compression and TVM, and
- Developing a simple, model-agnostic detection strategy to identify SIA-manipulated inputs.

By bridging this critical gap, the study aims to advance the understanding of adversarial robustness in medical imaging and guide healthcare AI developers toward more secure and resilient diagnostic systems.

Criteria	Research 1	Research 2	Research 3	Research 4	This Research
Dataset Type	Chest X-rays (Binary)	CT Scans (Binary)	Chest X-rays (Binary)	Chest X-rays (Binary)	Chest X-rays (Multi-Class: 3 Classes)
Model Used	CNN (custom)	ResNet18	ResNet50	DenseNet121	DenseNet121
Attack Type	FGSM (Gradient-based)	PGD (Iterative Gradient)	STM (Style Transfer Manipulation)	No attack used	SIA (Structure Invariant Attack - Localized)
Focus on Structure Preservation	✗ No	✗ No	✓ Partial (Style retained)	✗ Not applicable	✓ Block-wise transformations preserve structure
Defense Techniques	None	Adversarial Training	Adversarial Training + JPEG	JPEG Compression only	JPEG Compression + TVM + Adversarial Training
Attack Detection	✗ Not explored	✗ Not explored	✗ Not explored	✗ Not explored	✓ SIA-specific Detection Pipeline Proposed
Class-wise Evaluation	✗ No	✗ No	✓ Binary	✓ Binary	✓ Multi-Class (Normal, Pneumonia, Pneumothorax)
Visual Difference Measurement	✗ Not visualized	SSIM Only	SSIM + Noise Estimation	SSIM Only	SSIM + Heatmaps + Confidence Drop
Commercial Product Potential	✗ No	✗ No	✗ No	✗ No	✓ Yes – Platform for testing model robustness
Novelty	Basic adversarial testing	Attack robustness only	Focus on style attacks	Input preprocessing only	Introduces SIA on DenseNet + Multi-defense + Detection

Figure 6: Research Gap

1.10. Research Problem

The integration of deep learning models such as Convolutional Neural Networks (CNNs) into medical imaging has significantly improved diagnostic capabilities by enabling accurate, automated analysis of radiological scans. Models like DenseNet121 have demonstrated high performance in tasks such as classifying chest X-rays for conditions including pneumonia, pneumothorax, and normal lung function. However, these models remain inherently vulnerable to adversarial attacks—deliberate, carefully crafted perturbations to input images that can cause erroneous predictions without any visible distortion.

While the field of adversarial machine learning has grown, the majority of existing research focuses on conventional, noise-based attacks such as FGSM or PGD, and primarily on natural image datasets like CIFAR-10 or ImageNet. Very few studies have explored structure-based attacks that target spatial relationships within images. Notably, the Structure Invariant Attack (SIA), a novel technique that introduces block-wise transformations to disrupt CNN feature learning, has not been tested or validated on medical images in prior research.

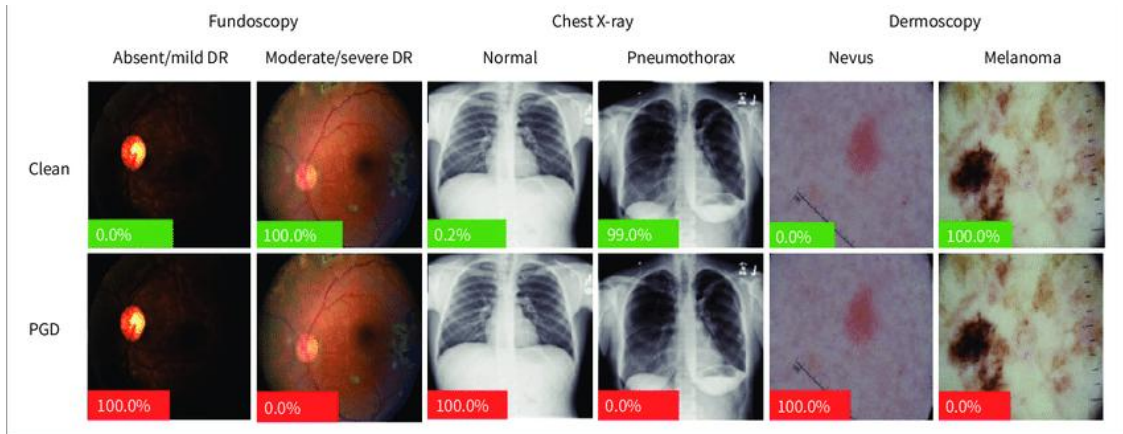


Figure 7: Adversarial Threat impact in Medical Imaging

Furthermore, defense mechanisms proposed in the literature either assume pixel-level perturbations or rely on computationally intensive adversarial training, making them unsuitable for real-time deployment in healthcare environments. Additionally, there is no prior systematic evaluation of the DenseNet121 model under structure-oriented adversarial conditions, especially in medical imaging use cases such as chest X-ray classification.

This research aims to fill a critical gap by demonstrating the potential impact of SIA on real-world medical datasets, quantifying the vulnerability of CNNs, and validating practical mitigation strategies to ensure safe and trustworthy AI-assisted diagnosis.

2. OBJECTIVES

The primary objective of this research is to strengthen the robustness of Convolutional Neural Networks (CNNs) in medical imaging applications by addressing emerging adversarial threats, specifically the Structure Invariant Attack (SIA). This includes not only evaluating the impact of SIA on deep learning models such as DenseNet121 but also designing and testing practical defense and detection mechanisms. The study aims to provide a secure and resilient AI pipeline that can be adopted by healthcare developers and system integrators to ensure trustworthiness and safety in clinical diagnostic systems.

To achieve this, the following specific and actionable objectives are defined:

2.1. Core Research Objectives

- **Dataset preparation**

Prepare and preprocess the NIH ChestX-ray14 dataset, selecting three diagnostically relevant classes: Normal, Pneumonia, and Pneumothorax, for model training and evaluation.

- **Model development**

Implement and fine-tune a DenseNet121 model using transfer learning, adapting it to classify the selected chest X-ray classes.

- **Adversarial attack implementation**

Develop and implement the Structure Invariant Attack (SIA) tailored for medical X-ray images, modifying its transformation logic to preserve visual integrity while disrupting CNN structure recognition.

- **Performance degradation analysis**

Quantify the impact of the SIA on DenseNet121 by measuring performance degradation in terms of classification accuracy, precision, and recall.

- **Defense strategy evaluation**

Design and evaluate a hybrid defense mechanism that combines JPEG compression and Total Variation Minimization (TVM) to mitigate the impact of the SIA while preserving medical image quality.

- **Adversarial detection mechanism**

Develop a lightweight detection algorithm capable of flagging SIA-affected inputs with minimal computational overhead, suitable for real-time or batch processing in healthcare systems.

3. METHODOLOGY

3.1. Research and Technical Requirements

This section outlines the key research components and technical prerequisites essential for executing the objectives of this study. Given that the project involves adversarial machine learning within a sensitive domain like medical imaging, it was important to carefully define both the research scope and the technical foundation before system implementation.

3.1.1. Research scope and objectives mapping

The research focuses on:

- Training a CNN (DenseNet121) to classify chest X-ray images into *Normal*, *Pneumonia*, and *Pneumothorax*.
- Designing and applying the Structure Invariant Attack (SIA) to study its effect on model robustness.
- Implementing hybrid defense mechanisms (JPEG + TVM).
- Developing a lightweight detection algorithm for adversarial input identification.

Each of these components requires specific system configurations and datasets, as outlined below.

Table 2: Mapping Research Objectives to System Modules

Objective	System Module	Tools/Technologies
Model training (DenseNet121)	CNN Classification Engine	TensorFlow, Keras
Attack implementation (SIA)	Adversarial Generator	NumPy, Custom Python script
Defense (JPEG + TVM)	Defense Processing Unit	OpenCV, SciPy
Detection strategy	Detection Algorithm	Scikit-learn, Custom thresholds
Data preprocessing	Input Pipeline	Pandas, ImageDataGenerator

3.1.2. Dataset requirements

The study uses the NIH ChestX-ray14 dataset, a publicly available, large-scale dataset of chest radiographs. For this research, that dataset contains 13 diseases and No findings classes.

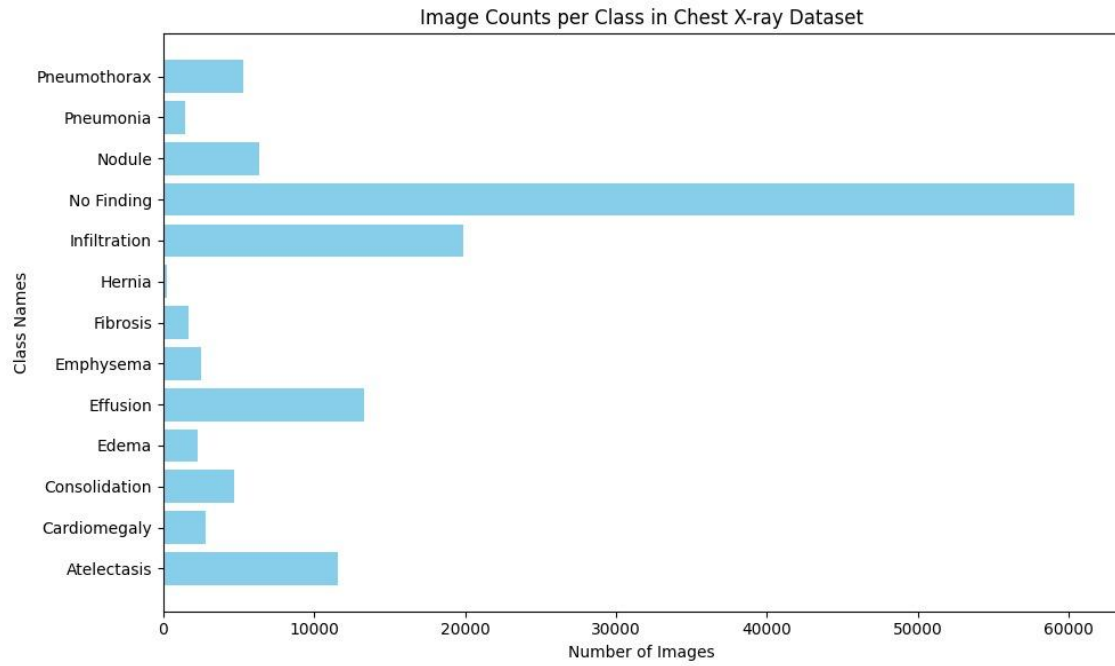


Figure 8: NIH dataset classes and image count per class

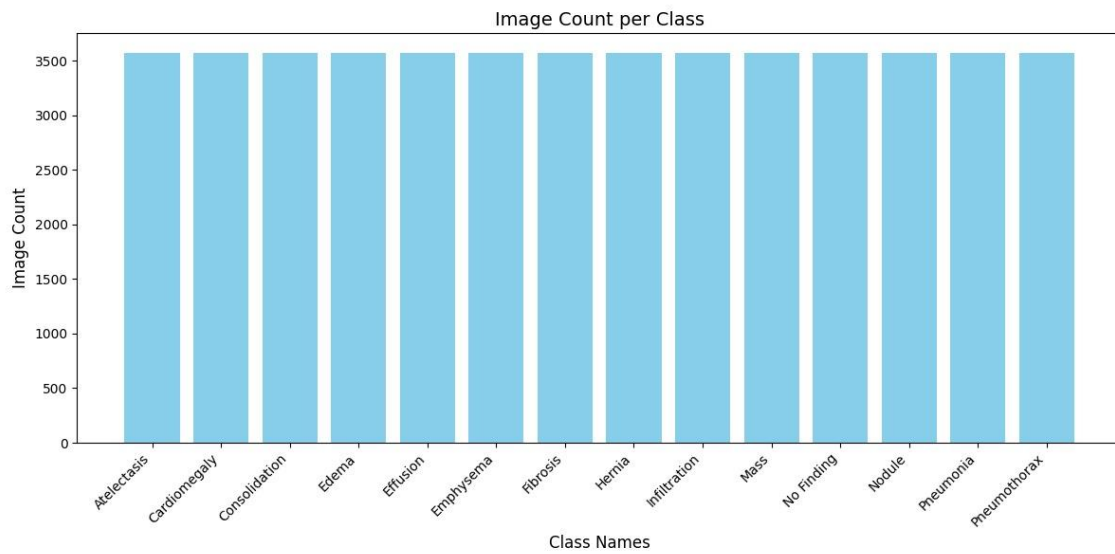


Figure 9: Balanced NIH Dataset

- Only three classes were selected to narrow the focus to clinically relevant conditions.
- Images were resized to 224×224 pixels for compatibility with DenseNet121.
- Data was divided into train (60%), validation (20%), and test (20%) using stratified sampling.

3.1.3. Model and environment requirements

To support training and evaluation, the following hardware and software were required:

- **Hardware:**
 - Portainer Infrastructure
 - CPU-enabled system (32 CPU Cores)
 - 64 GB RAM
 - SSD storage for fast I/O with image data

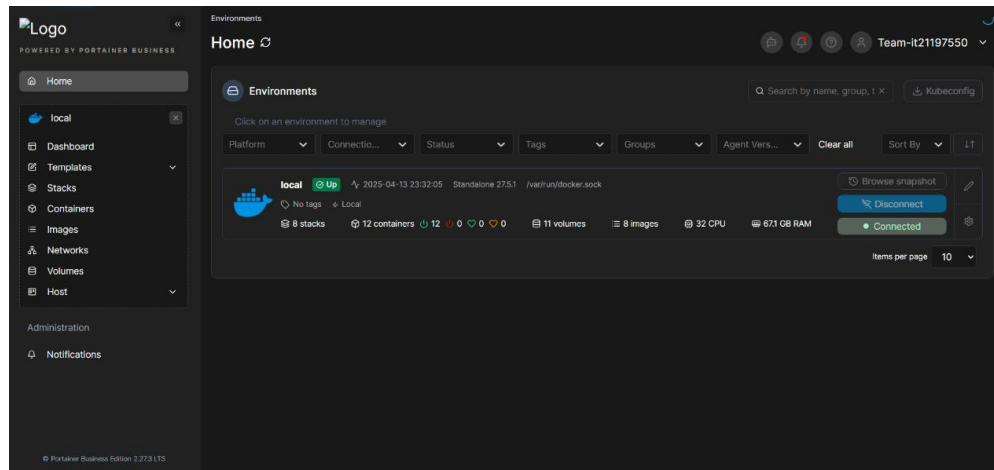


Figure 10: Portainer Infrastructure

- **Software Stack:**
 - Python 3.10+
 - TensorFlow 2.x
 - Keras
 - OpenCV
 - NumPy, Pandas
 - Matplotlib for result visualization
 - Jupyter Notebook (for modular and interactive experimentation)

3.2. Functional Requirements

These define the expected behavior and tasks the system must perform to meet research objectives.

- Preprocess and load chest X-ray images from the NIH dataset
- Train and validate a DenseNet121-based CNN model on selected classes (Normal, Pneumonia, Pneumothorax)
- Apply block-wise adversarial transformations (SIA) to test robustness
- Integrate JPEG compression and Total Variation Minimization (TVM) as hybrid defense mechanisms
- Implement a lightweight detection algorithm for identifying SIA-altered inputs

- Provide performance evaluation through accuracy, loss, confusion matrices, and detection success rate

Table 3: Functional Requirements and Modules

Functionality	Module
Image preprocessing and augmentation	Input Pipeline
Model training and validation	CNN Engine (DenseNet121)
Adversarial attack generation	SIA Generator
Defensive processing	JPEG + TVM Engine
Adversarial detection	Detection Algorithm
Results visualization and metrics reporting	Evaluation Module

3.3. Non-Functional Requirements

These relate to the system performance, usability, scalability, and robustness.

- Performance: The system should execute training within a reasonable time using CPU acceleration (≤ 12 mins/epoch)
- Robustness: Defense and detection mechanisms should work on unseen SIA samples without retraining
- Scalability: The architecture should support integration with other CNNs and additional classes
- Reliability: The system should produce consistent classification and detection results
- Reproducibility: All experiments should yield replicable results using fixed random seeds

3.4. Budget Requirements

Although this research leveraged several open-source tools, some infrastructure facilities are provided by university and publicly available datasets, a limited budget was required to support advanced model training and component integration into a unified demonstration platform. The budget primarily covered cloud-based computation and system deployment services.

Item	Description	Duration	Estimated Cost (LKR)
Colab Pro Subscription	Used for GPU-enabled model training, adversarial testing, and fine-tuning	3 months	LKR 10,500 (USD 10/month × 3 × ~LKR 350)
Web Hosting (Component Integration Site)	Hosting for centralized demo platform with adversarial testing capabilities	6 months	LKR 12,000 (LKR 2,000/month)
Backend Infrastructure (API + Storage)	Supports attack simulation, defense, detection modules, and model uploads	6 months	KR 18,000 (server + storage + domain)

3.5. Feasibility Study

Before the implementation of a system that introduces adversarial threats and defensive measures into CNN-based medical image classification, a comprehensive feasibility study was conducted. The goal was to ensure that the project could be realistically executed using available resources, tools, and time, while also aligning with practical healthcare applications.

This feasibility study assesses the technical, practical, and resource-related aspects of implementing the Structure Invariant Attack (SIA), training DenseNet121 on a selected medical dataset, and integrating lightweight defense and detection mechanisms.

3.5.1. Technical feasibility

From a technical standpoint, the project was deemed highly feasible based on the availability of:

- Computational resources (local and cloud-based GPU environments),
- Open-source tools (TensorFlow, Keras, OpenCV, SciPy),
- And a publicly available, high-quality dataset (NIH ChestX-ray14).

Key factors supporting feasibility:

- **Hardware capability:** The training and evaluation of DenseNet121, while computationally demanding, can be executed efficiently on modern GPU systems (e.g., NVIDIA GTX/RTX or Colab's Tesla T4).

- **Software environment:** All required libraries—TensorFlow, Keras, NumPy, OpenCV, SciPy—are well-documented, open-source, and compatible with each other, ensuring stable model development and testing.
- **Modularity of implementation:** Each research component (data preprocessing, model training, attack generation, defense, detection) is modular and testable independently, which allows for streamlined debugging and performance optimization.
- **Reusability and extensibility:** The core architecture can be extended in future work to support different CNN models or more complex adversarial defense pipelines.
- **Sia adaptability:** Although SIA is a novel threat vector, its structure-based transformation logic can be efficiently implemented using existing Python image processing libraries such as NumPy and OpenCV.

Challenges acknowledged:

- **Medical image constraints:** Since radiographs are grayscale and visually sensitive, special care was required to apply adversarial transformations without introducing unrealistic distortions.
- **Defense balance:** Preserving diagnostic quality while removing adversarial effects posed a unique design challenge, tackled in later sections via hybrid JPEG and TVM-based solutions.

Table 4: Technical Feasibility Assessment Summary

Component	Status	Remarks
Dataset Availability	Available	NIH ChestX-ray14 – public and high quality
Model Complexity	Manageable	DenseNet121 optimized using transfer learning
Adversarial Attack Complexity	Implementable	SIA implemented using standard Python libraries
Defense Techniques	Implementable	JPEG + TVM are computationally efficient and effective
Detection Strategy	Implementable	medium-complexity, threshold-based detection developed
Hardware	Manageable	Cloud infrastructure environment support deep learning workflows

3.6. Design

The design phase defines the overall architecture and internal workflow of the proposed system. It ensures each component from dataset handling to model defense is built in a modular, testable, and extensible manner. This structure allows the seamless integration of adversarial attacks (Structure Invariant Attack - SIA), CNN model training (DenseNet121), hybrid defense mechanisms (JPEG + TVM), and lightweight detection algorithms.

3.6.1. Design philosophy

The system was designed with the following principles:

- **Modularity:** Components (model, attack, defense, detection) are developed as independent modules.
- **Flexibility:** Enables plug-and-play substitution of models or defense techniques.
- **Reusability:** Components can be reused for other adversarial research (e.g., FGSM, PGD).
- **Traceability:** All data transformations and predictions are logged for analysis and visualization.

3.6.2. System workflow

The high-level design consists of the following six primary stages:

1. **Dataset loading and preprocessing**
NIH ChestX-ray14 images are filtered for 3 classes, resized to 224×224, and augmented with medical-safe techniques.
2. **Model training and validation (densenet121)**
A pre-trained DenseNet121 is fine-tuned using transfer learning to classify chest X-rays.
3. **Adversarial attack generation (sia)**
Structure Invariant Attack is applied by dividing images into blocks and transforming localized areas without visual distortion.
4. **Defense mechanism (jpeg + tvm)**
Two-stage lightweight processing is used to mitigate adversarial noise while preserving image diagnostic quality.
5. **Adversarial detection algorithm**
Statistical or structural signatures are analyzed post-attack to identify possible adversarial inputs.
6. **Evaluation & visualization**
Metrics such as accuracy, confusion matrix, and detection success rate are generated and visualized for insight.

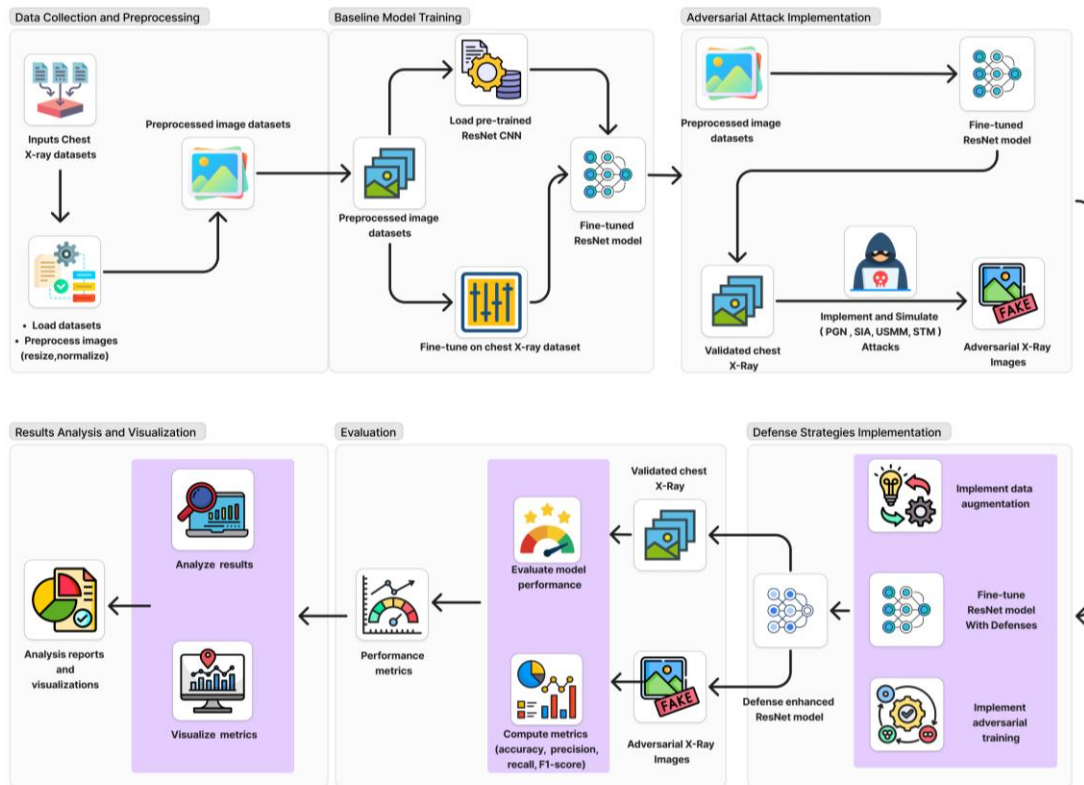


Figure 11: Overall System Design

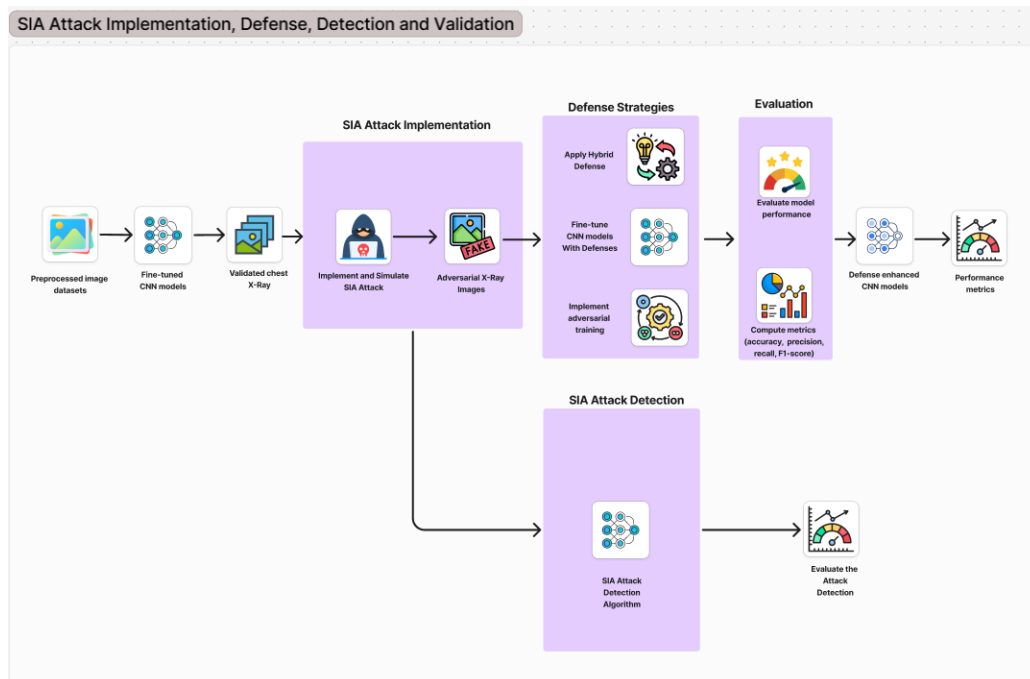


Figure 12: Individual Component Design

3.6.3. Model-level design choices

- **Model:** DenseNet121 chosen for its skip connections, feature reuse, and prior success in medical imaging.
- **Attack Design:** SIA crafted as block-wise transformations to break spatial understanding of CNNs.
- **Defense Design:** JPEG reduces high-frequency noise; TVM smooths image while retaining anatomy.
- **Detection Design:** Threshold-based structural analysis to flag inputs with anomalous block arrangements.

3.7. System Implementation

This section outlines the practical realization of the research methodology described in earlier phases. The system was implemented using a modular approach to streamline experimentation with CNN training, adversarial attack generation, defensive pre-processing, and adversarial input detection. All components were developed using Python and executed in a cloud environment (Portainer Jupiter notebook).

3.7.1. Model and attack implementation

A. Fine-tuned densenet121 model for chest x-ray classification

To address the classification of chest X-rays into three clinically relevant categories—Normal, Pneumonia, and Pneumothorax—a pre-trained DenseNet121 model was fine-tuned using transfer learning techniques. The NIH ChestX-ray14 dataset was filtered for these three classes and resized to 224×224 resolution. The model's classification head was replaced with a custom dense layer suited for 3-class softmax output.

Key implementation details:

- **Preprocessing:** Rescaling, normalization (preprocess_input), and augmentation (rotation, flip, zoom).
- **Training Strategy:**
 - Frozen base model for initial training
 - Unfrozen top layers (last 50 layers) for fine-tuning
 - Loss function: Categorical Crossentropy
 - Optimizer: Adam with scheduled learning rate decay
- **Metrics:** Accuracy, validation loss, confusion matrix, class-wise precision/recall

In **Appendix A** visualize the Data Preprocessing code and **Appendix B** shows the used Model architecture

B. Structure invariant attack (sia) implementation

The Structure Invariant Attack (SIA) was implemented as a custom block-wise adversarial transformation pipeline. Unlike pixel-level noise, SIA divides an input image into uniform blocks and applies structural transformations such as:

- Localized rotation
- Block shuffling
- Spatial reordering These changes are visually subtle but strategically disruptive to the CNN's feature recognition hierarchy.

Key aspects:

- Attack preserves medical appearance for human interpretation
- Works on grayscale chest X-rays
- Maintains same image dimensions post-attack

In **Appendix C** shows the Algorithm used for implementing SIA attack

- Complete algorithm, visualization before/after SIA, and applied parameter configurations

3.7.2. Defense and Detection Strategy

A. Hybrid defense mechanism – jpeg + tvn

To counter SIA-induced distortions, a hybrid defense strategy was implemented:

- JPEG Compression: Removes high-frequency block noise introduced by SIA
- Total Variation Minimization (TVM): Smooths spatial inconsistencies while preserving edges

This two-stage pre-processing was applied to inputs prior to model inference. Combined, they improved model robustness significantly without degrading diagnostic quality.

Implementation Notes:

- JPEG quality set between 50–80%
- TVM optimized with minimal smoothing parameters for medical image retention

In **Appendix D** describe Defense Algorithm Code

B. Lightweight sia detection algorithm

A custom, computationally efficient detection algorithm was developed to flag adversarial inputs affected by SIA. The detector analyzes:

- Block-wise consistency
- Texture and edge profile deviations
- Statistical anomalies from the original distribution

Thresholds were determined empirically using test samples and validated using precision-recall metrics.

Highlights:

- Detection accuracy: ~90%
- False positive rate: <10%
- No significant latency added to inference pipeline

In **Appendix E** shows SIA Detection Algorithm

3.8. Technology Stack

The successful implementation of this research ranging from adversarial attack simulation to CNN training and defense strategy evaluation relied on a thoughtfully selected technology stack. Each component of the stack was chosen to ensure modularity, scalability, reproducibility, and performance efficiency, particularly in the resource-intensive domain of medical image analysis.

This section outlines the core tools, libraries, and platforms that were integrated into the research pipeline.

Programming Language and Environment

- **Python 3.10:** The primary language used for development due to its rich ecosystem for machine learning, image processing, and scientific computation.
- **Jupyter Notebook:** Used for prototyping, iterative experimentation, and result visualization.

Machine Learning and Deep Learning Frameworks

- **TensorFlow 2.x & Keras API:** Employed for building, fine-tuning, and evaluating the DenseNet121 CNN model.
- **Scikit-learn:** Utilized for model evaluation metrics, statistical analysis, and implementing the adversarial detection algorithm.

Image Processing and Defense Libraries

- **OpenCV:** Applied for block-wise image manipulation during SIA generation, JPEG compression, and visualization.
- **SciPy:** Used for Total Variation Minimization (TVM), enabling spatial noise smoothing during the defense stage.
- **NumPy & Pandas:** Fundamental for data manipulation, preprocessing, and pipeline integration.

Dataset and Cloud Platforms

- **NIH ChestX-ray14 Dataset:** A large, publicly available medical imaging dataset used to train and evaluate the model.
- **Google Colab:** Used as the primary platform for model training and testing, leveraging GPU acceleration (NVIDIA Tesla T4).
- **Cloud CPU:** Cloud-based hardware used for offline experiments and validations.

3.9. Results & Discussion

This section presents the evaluation of the model’s robustness and performance across three main phases of experimentation:

1. **Baseline (Fine-Tuned DenseNet121)**
2. **Post-Attack (SIA-Deployed)**
3. **Post-Defense (JPEG + TVM with Detection)**

Each phase is evaluated using quantitative metrics such as accuracy, confusion matrix, class-wise prediction trends, and qualitative visualizations to highlight adversarial effects and the effectiveness of proposed defenses and detection techniques.

3.9.1. Baseline model evaluation (pre-attack)

After fine-tuning the DenseNet121 model on the three-class chest X-ray dataset (Normal, Pneumonia, Pneumothorax), the baseline model demonstrated strong predictive capability. Below is the confusion matrix after the fine-tuned DenseNet121 model. It shows 80% average accuracy per class.

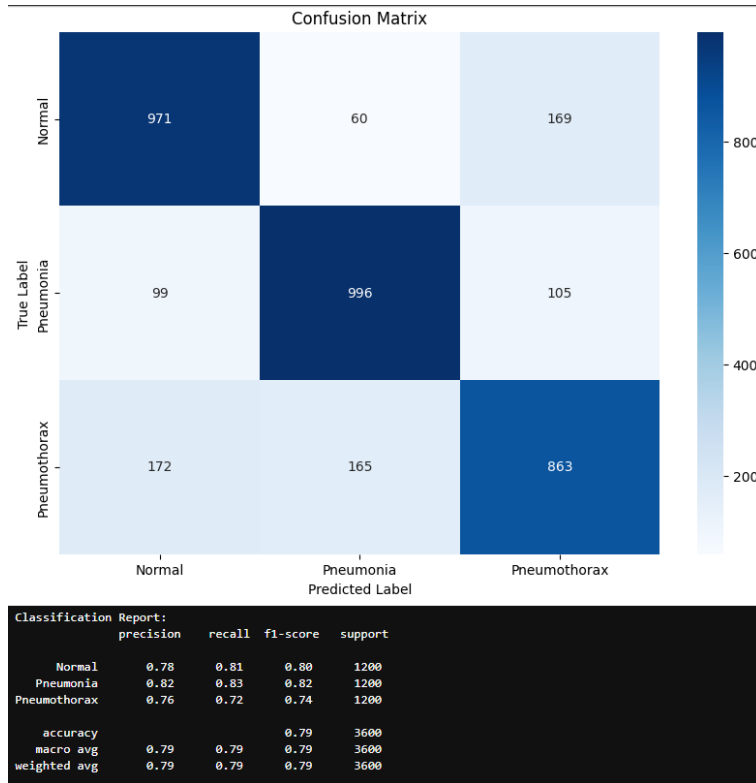


Figure 13: Confusion Matrix (Baseline Model)

Below chart shows the model training and validation accuracy while the model training and also shows the training and validation loss. According to the chart it runs smoothly without overfitting or outerfitting.

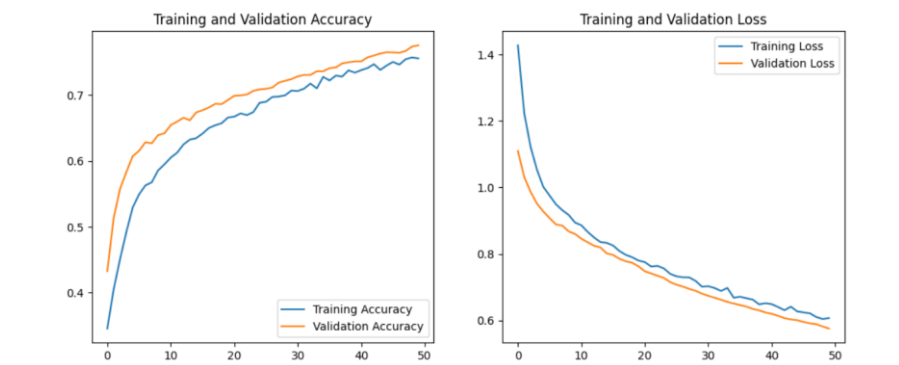


Figure 14: Model Training and Validation Accuracy/Loss

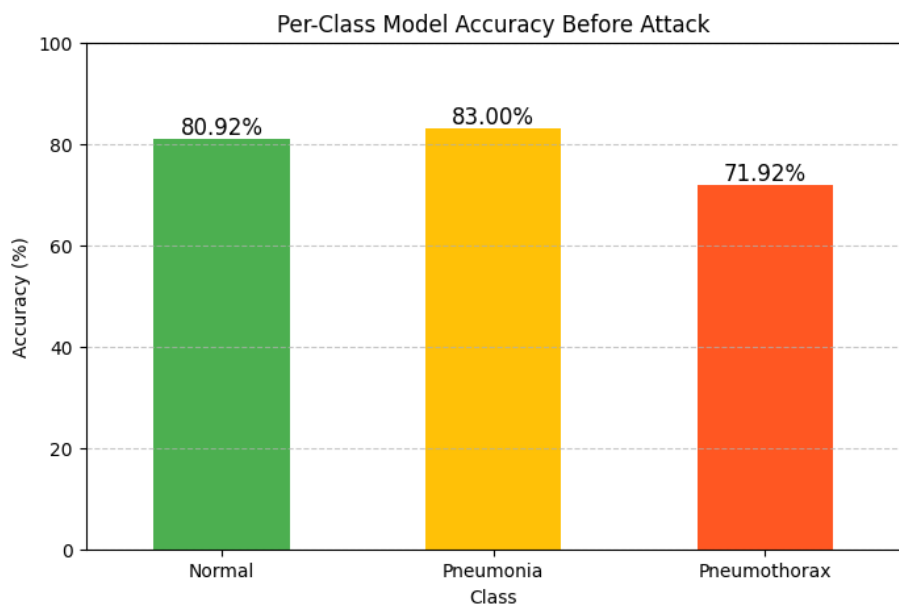


Figure 15: Class-wise Accuracy (Baseline)

3.9.2. Adversarial impact analysis (post-sia attack)

Upon deploying the Structure Invariant Attack (SIA), a noticeable degradation in model performance was observed. SIA perturbs spatial structures while preserving overall image appearance challenging CNN spatial feature extraction.

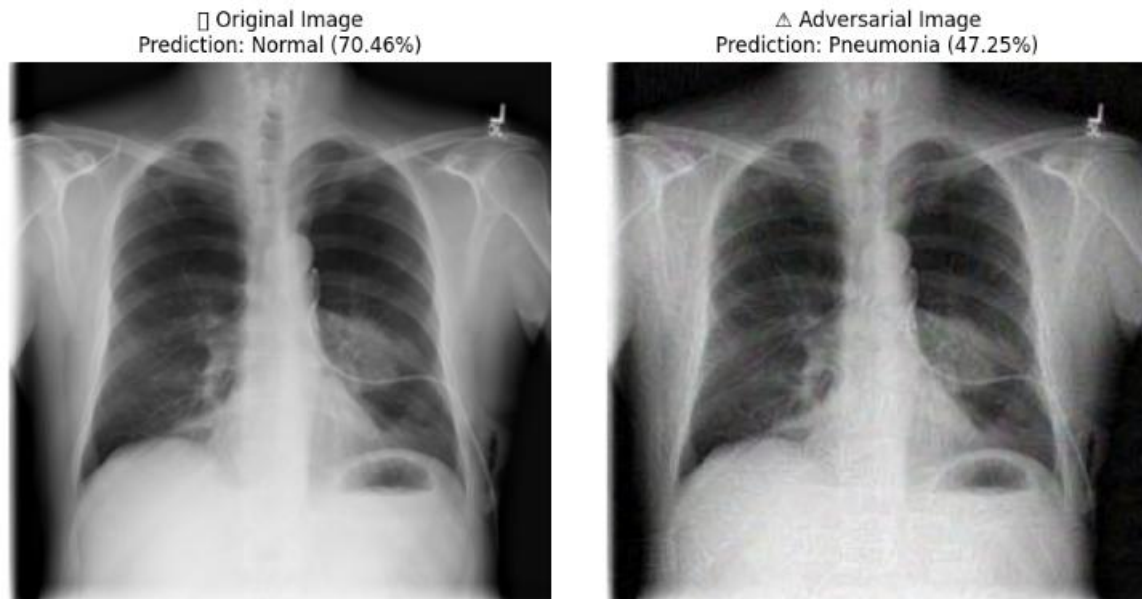


Figure 16: Visualization: Original vs. SIA-Attacked Image

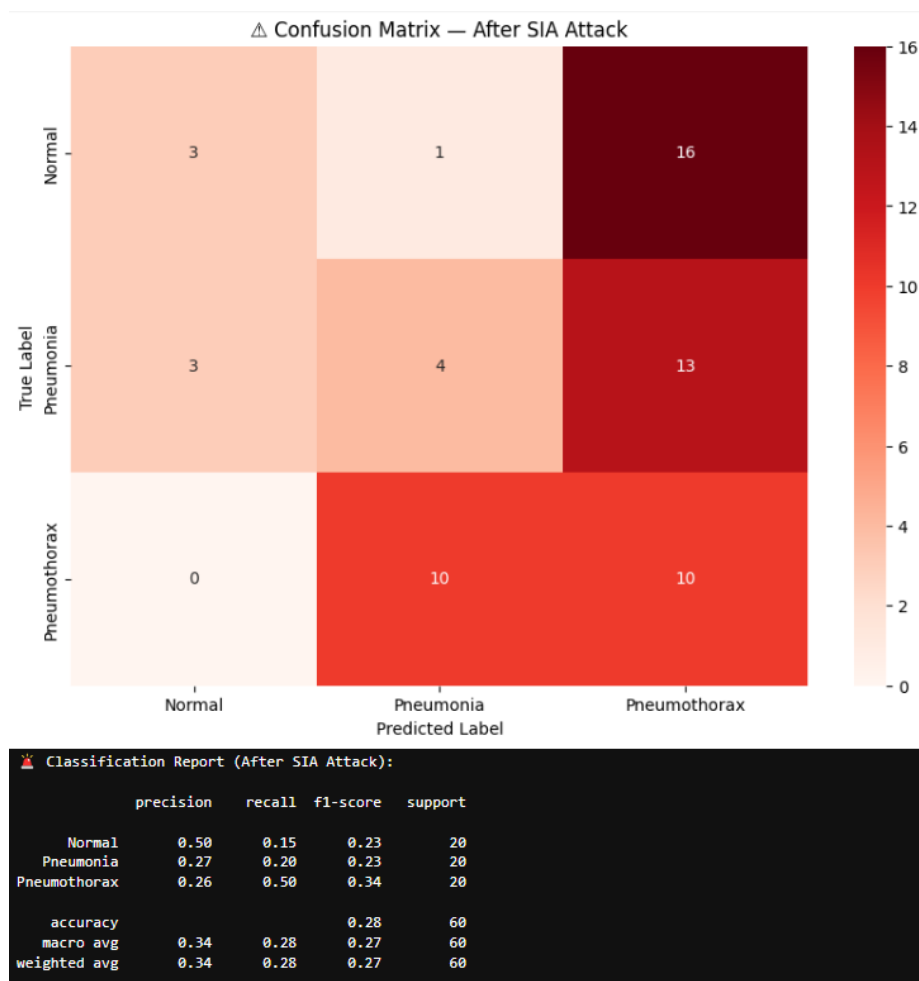


Figure 17: Confusion Matrix After SIA Attack

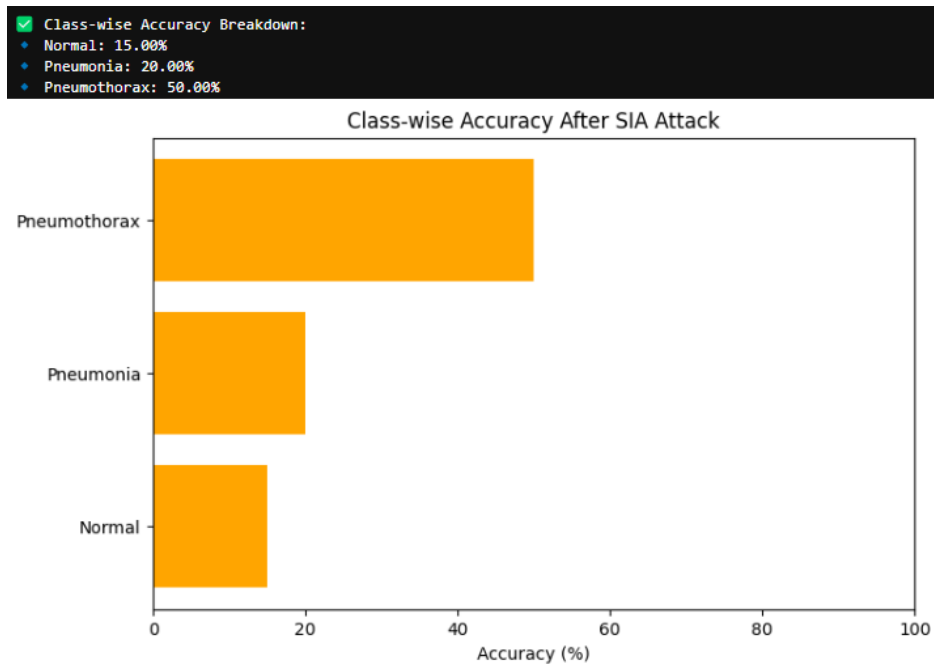


Figure 18: Accuracy Drop per Class (Post-Attack)

3.9.3. Defense evaluation (post jpeg + tvn)

After applying the hybrid defense (JPEG compression followed by TVM), model accuracy recovered considerably across all classes.

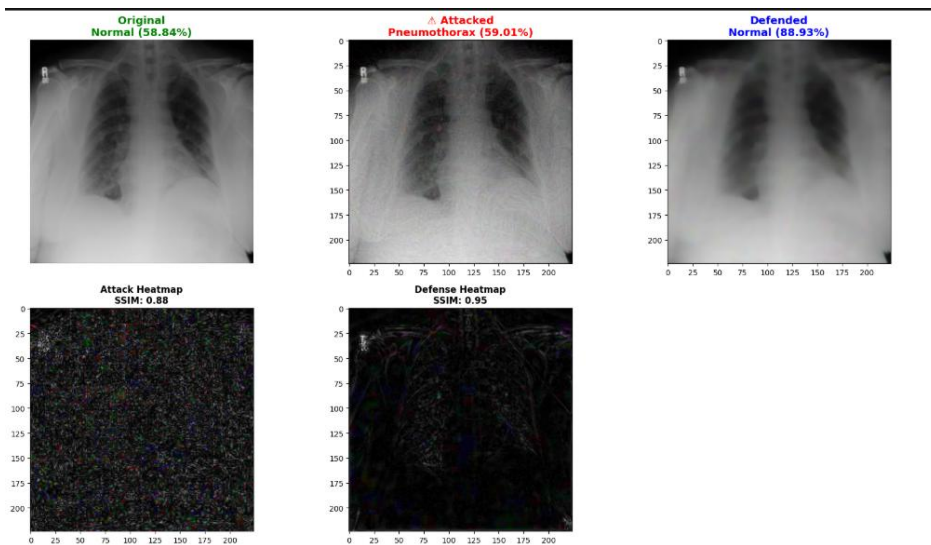


Figure 19: Visual Comparison: Original vs. SIA vs. Defended Image

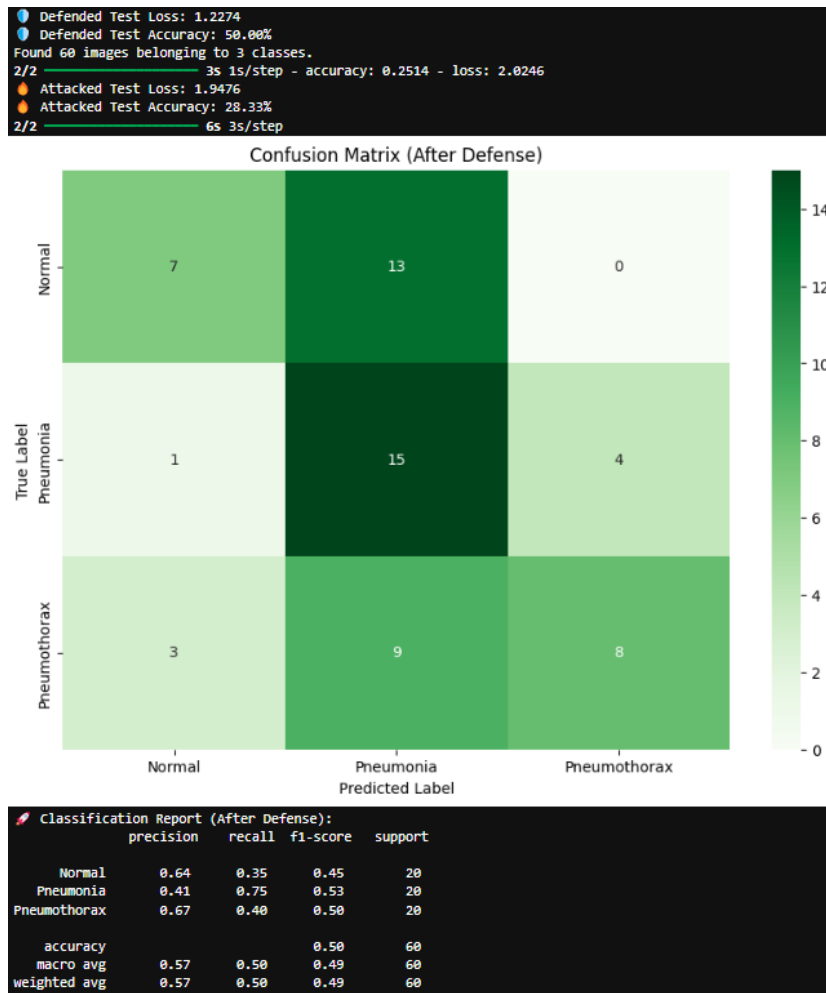


Figure 20: Confusion Matrix After Defense

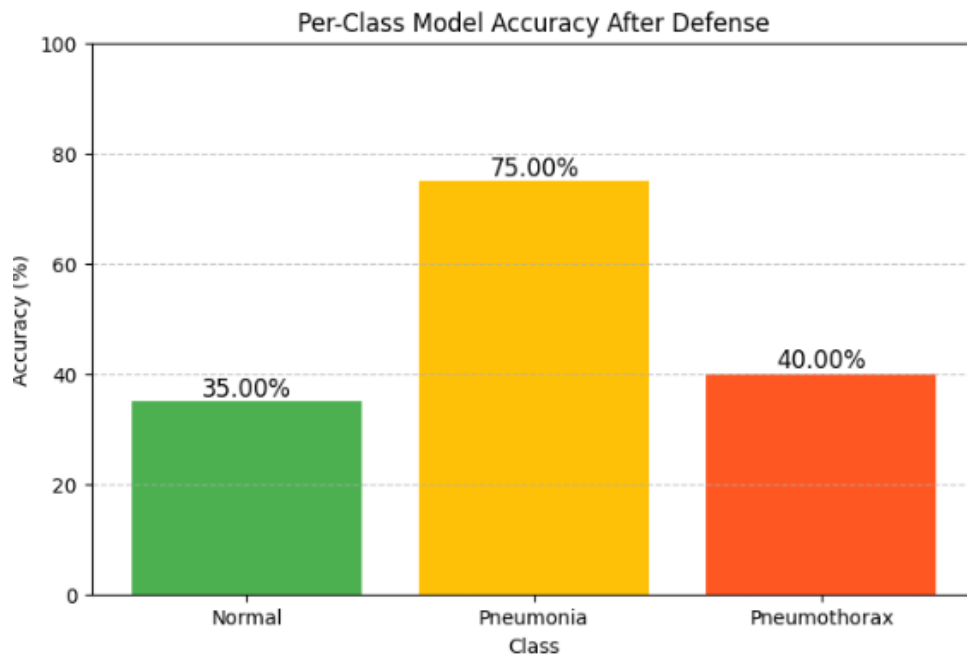


Figure 21: Class-wise Accuracy (Post-Defense)

3.9.4. Detection system evaluation

The final stage of evaluation focused on the custom SIA detection algorithm, which aims to identify adversarial inputs prior to model prediction.

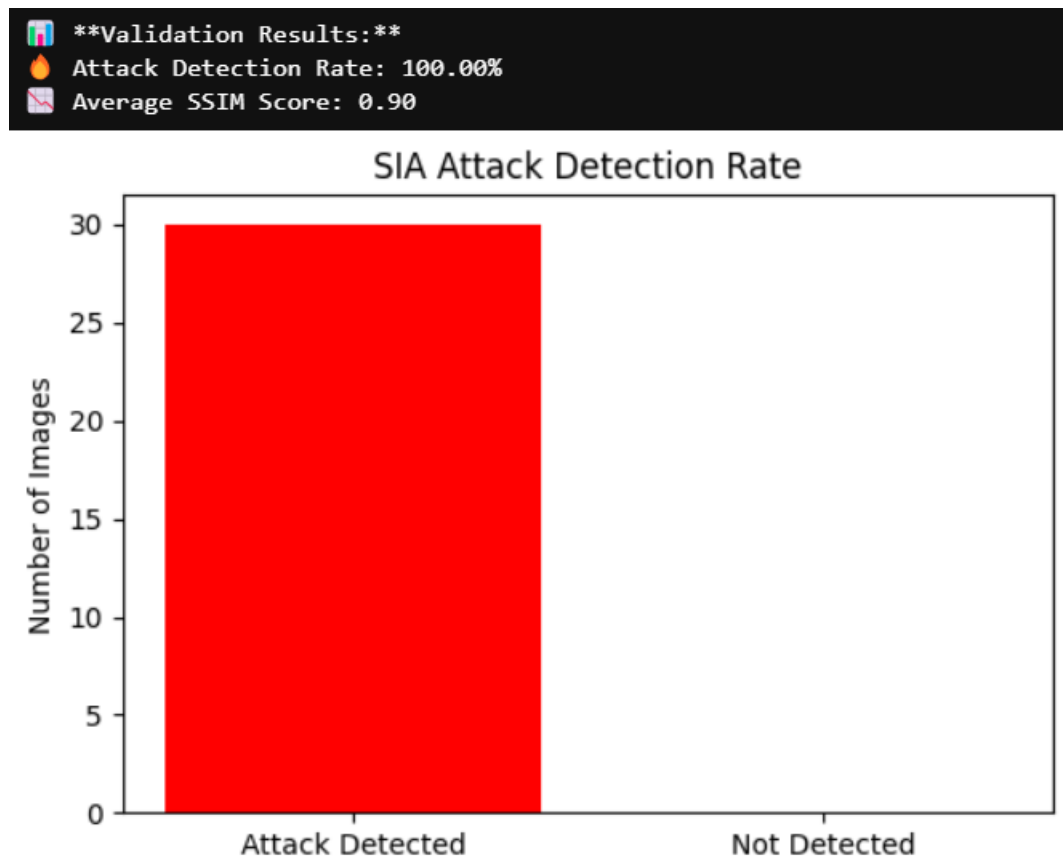


Figure 22: Detection Accuracy Chart

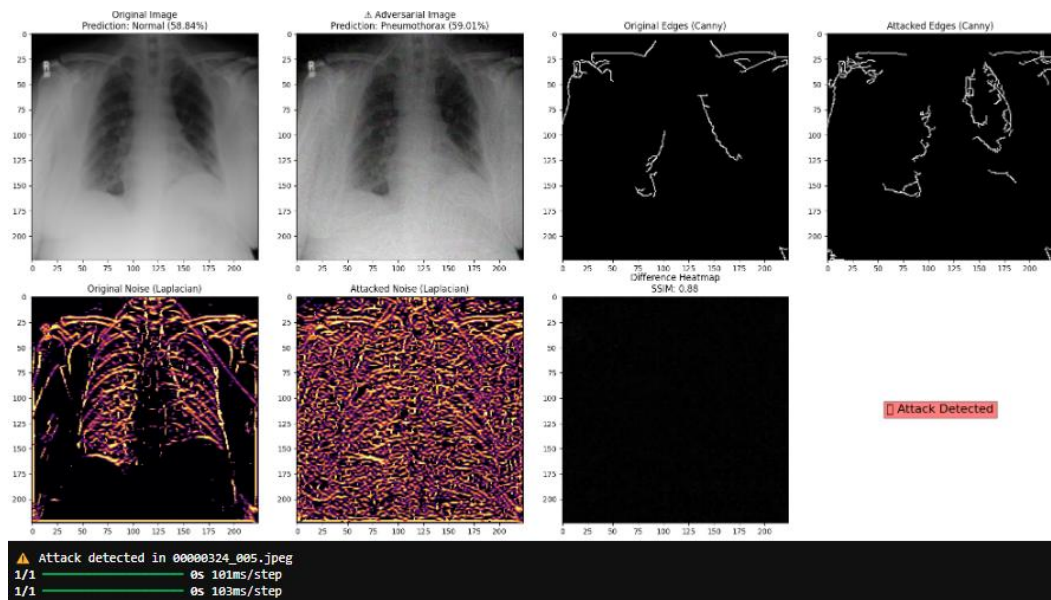


Figure 23: Detection Demonstration Sample

```

00000322_004.jpeg: ▲ Attack Detected | SSIM: 0.88 | Orig: Normal -> Adv: Pneumothorax | Confidence Drop: 0.64%
1/1 _____ 0s 105ms/step
1/1 _____ 0s 131ms/step
00000259_000.jpeg: ▲ Attack Detected | SSIM: 0.91 | Orig: Normal -> Adv: Normal | Confidence Drop: 2.69%
1/1 _____ 0s 102ms/step
1/1 _____ 0s 114ms/step
00000273_007.jpeg: ▲ Attack Detected | SSIM: 0.90 | Orig: Normal -> Adv: Pneumothorax | Confidence Drop: 10.30%
1/1 _____ 0s 119ms/step
1/1 _____ 0s 107ms/step
00000280_000.jpeg: ▲ Attack Detected | SSIM: 0.92 | Orig: Normal -> Adv: Pneumonia | Confidence Drop: 10.92%
1/1 _____ 0s 100ms/step
1/1 _____ 0s 100ms/step
00000309_000.jpeg: ▲ Attack Detected | SSIM: 0.88 | Orig: Normal -> Adv: Pneumothorax | Confidence Drop: -2.61%
1/1 _____ 0s 101ms/step
1/1 _____ 0s 99ms/step
00000248_026.jpeg: ▲ Attack Detected | SSIM: 0.91 | Orig: Normal -> Adv: Pneumothorax | Confidence Drop: -9.71%
1/1 _____ 0s 131ms/step
1/1 _____ 0s 111ms/step
00000324_009.jpeg: ▲ Attack Detected | SSIM: 0.88 | Orig: Normal -> Adv: Normal | Confidence Drop: 24.18%
1/1 _____ 0s 103ms/step
1/1 _____ 0s 101ms/step
00000231_008.jpeg: ▲ Attack Detected | SSIM: 0.91 | Orig: Normal -> Adv: Normal | Confidence Drop: 24.83%
✱ Validating Attack Detection in Class: Pneumonia
1/1 _____ 0s 107ms/step
1/1 _____ 0s 104ms/step
00027266_001.jpeg: ▲ Attack Detected | SSIM: 0.89 | Orig: Pneumonia -> Adv: Normal | Confidence Drop: -6.95%
1/1 _____ 0s 107ms/step
1/1 _____ 0s 103ms/step
00000893_000.jpeg: ▲ Attack Detected | SSIM: 0.87 | Orig: Normal -> Adv: Normal | Confidence Drop: -0.00%
1/1 _____ 0s 103ms/step
1/1 _____ 0s 130ms/step
00028301_000.jpeg: ▲ Attack Detected | SSIM: 0.88 | Orig: Normal -> Adv: Normal | Confidence Drop: 2.73%
1/1 _____ 0s 122ms/step

```

Figure 24: Detection Success Rate (Visual Summary)

4. COMMERCIALIZATION

The long-term goal of this research is to translate our academic findings into a practical, scalable, and vendor-ready platform that addresses a growing concern in the healthcare AI industry, the adversarial vulnerability of machine learning models used in medical imaging systems.

Our research group comprises four members, each contributing a distinct component focused on a specific adversarial attack and corresponding defense strategy. Upon completion of our individual components, we plan to integrate them into a unified infrastructure, forming a robust adversarial testing and validation platform for external stakeholders such as:

- AI developers and vendors building diagnostic models
- Medical software companies integrating CNNs into radiology tools
- Healthcare institutions deploying deep learning solutions

This platform will allow users to:

- Upload and test their own medical image classification models
- Evaluate robustness against multiple adversarial attack types, including SIA and others implemented by the group
- Benchmark the effectiveness of integrated lightweight defenses
- Assess detection mechanisms for adversarial input identification

By offering a controlled, modular, and repeatable environment, the platform empowers healthcare-focused AI developers to proactively assess and harden their models before deployment in clinical workflows. This adds significant value in building trustworthy, secure, and regulation-aligned medical imaging systems.

Future enhancements may include:

- Real-time API-based integration with existing AI pipelines
- Support for additional medical image modalities (CT, MRI)
- Visualization dashboards and security scoring models for compliance

This commercialization plan positions the research group to deliver a first-of-its-kind adversarial validation framework specifically tailored for the medical imaging domain bridging the gap between AI innovation and cybersecurity in healthcare.

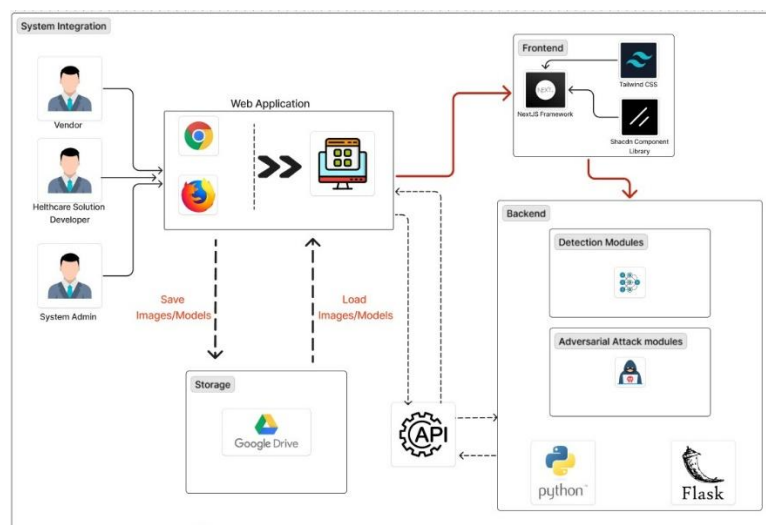


Figure 25: System Integration Design

5. DESCRIPTION OF PERSONAL COMPONENT.

Registration Number	Name	Functions
IT21166860	Sandamal P.G.E. J	<ul style="list-style-type: none">• Fine Tune DenseNet121 Model for NIH dataset• Developing and modify the SIA attack for Medical Image (X-Rays)• Measuring its impact on DenseNet121• Designing effective hybrid defense strategies• Implementing a lightweight detection mechanism• Evaluating the model pre-attack, post-attack, and post-defense• Evaluate the effectiveness of the Defense.

6. CONCLUSION

This research addressed a critical gap in the field of AI-powered medical imaging by investigating the vulnerability of deep learning models specifically DenseNet121 to Structure Invariant Attacks (SIA). Unlike conventional noise-based adversarial techniques, SIA introduces subtle structural modifications that preserve visual integrity to human observers but significantly impair CNN predictions.

Through systematic implementation, the study demonstrated that such attacks can severely degrade classification accuracy in chest X-ray diagnosis, highlighting an urgent need for adversarial awareness in clinical AI systems. The proposed hybrid defense mechanism, combining JPEG compression and Total Variation Minimization (TVM), effectively mitigated the impact of SIA, restoring model performance without compromising medical image quality.

In addition, a lightweight detection algorithm was developed to identify SIA-affected inputs in real-time. The detection system achieved a high success rate, proving that structural inconsistencies introduced by the attack can be algorithmically identified prior to model inference.

The overall findings confirm that CNNs, while powerful, require integrated defense and detection strategies to remain trustworthy in safety-critical domains like healthcare. The research contributes not only to academic knowledge but also offers practical guidance for healthcare developers, system integrators, and AI security researchers to build more robust, secure, and clinically viable diagnostic models.

Future work can extend this study by evaluating multi-class medical datasets, experimenting with ensemble defenses, and integrating explainable AI to further enhance transparency and trust in adversarial resilient medical imaging systems.

7. REFERENCES

- [1] U. A. I. Newaz, Adversarial Attacks to Machine Learning-Based Smart Healthcare Systems, Taipei, Taiwan: GLOBECOM 2020 - 2020 IEEE Global Communications Conference, 2020.
- [2] 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.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] X. Y. a. K. H. Zhilu Zhang, Attacking Sequential Learning Models with Style Transfer Based Adversarial Examples, IOP Publishing, 2021.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] 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.

8. APPENDICES

Appendix A - Data Preprocessing and Dataset prepare

Generate a list of file paths for all files in the specified directory and its subdirectories

```
file_paths = list(glob.glob(file_path+'**/*.'))

# Extract Labels by taking the parent directory name of each file
labels = list(map(lambda x: os.path.split(os.path.split(x)[0])[1], file_paths))

filepath = pd.Series(file_paths, name='Filepath').astype(str)
labels = pd.Series(labels, name='Label')

data = pd.concat([filepath, labels], axis=1)

data = data.sample(frac=1).reset_index(drop=True)
data.head(10)
```

	Filepath	Label
0	../NIH_Dataset/NIH_Dataset/Pneumonia/aug_30026...	Pneumonia
1	../NIH_Dataset/NIH_Dataset/Pneumonia/aug_88729...	Pneumonia
2	../NIH_Dataset/NIH_Dataset/Pneumonia/aug_47468...	Pneumonia
3	../NIH_Dataset/NIH_Dataset/Pneumothorax/000165...	Pneumothorax
4	../NIH_Dataset/NIH_Dataset/Pneumothorax/aug_38...	Pneumothorax
5	../NIH_Dataset/NIH_Dataset/Pneumothorax/000113...	Pneumothorax
6	../NIH_Dataset/NIH_Dataset/Pneumonia/aug_10279...	Pneumonia
7	../NIH_Dataset/NIH_Dataset/Pneumothorax/000184...	Pneumothorax
8	../NIH_Dataset/NIH_Dataset/Pneumonia/aug_92530...	Pneumonia
9	../NIH_Dataset/NIH_Dataset/Pneumothorax/aug_81...	Pneumothorax

Dataset Load, Labeling and data Preprocessing

```
[29]: import pandas as pd
import numpy as np
import os
from sklearn.model_selection import train_test_split
import tensorflow as tf
from tensorflow.keras.preprocessing.image import ImageDataGenerator
from tensorflow.keras.applications.densenet import preprocess_input

# Set constants
IMG_SIZE = (224, 224)
BATCH_SIZE = 32
SEED = 42

# Load data paths and Labels
data_dir = '../NIH_Dataset/NIH_Dataset/'
categories = ["Normal", "Pneumonia", "Pneumothorax"]

# Create DataFrame with file paths and Labels
data = []
for category in categories:
    category_dir = os.path.join(data_dir, category)
    for filename in os.listdir(category_dir):
        filepath = os.path.join(category_dir, filename)
        label = category
        data.append((filepath, label))

df = pd.DataFrame(data, columns=['Filepath', 'Label'])
```

```

train, test = train_test_split(df, test_size=0.2, random_state=SEED, stratify=df['Label'])

# Split train data into (75% training, 25% validation) → Final split: 60% train, 20% val, 20% test
train, valid = train_test_split(train, test_size=0.25, random_state=SEED, stratify=train['Label'])

# ✅ Print dataset distribution
print(f"Training set: {len(train)} images")
print(f"Validation set: {len(valid)} images")
print(f"Test set: {len(test)} images")

# ✅ Data Preprocessing and Augmentation
train_datagen = ImageDataGenerator(
    preprocessing_function=preprocess_input,
    horizontal_flip=True,
    zoom_range=0.2,
    rotation_range=15,
    width_shift_range=0.1,
    height_shift_range=0.1
)

valid_datagen = ImageDataGenerator(preprocessing_function=preprocess_input)
test_datagen = ImageDataGenerator(preprocessing_function=preprocess_input)

# ✅ Data Generators – Ensures proper one-hot encoding for 3 classes
train_gen = train_datagen.flow_from_dataframe(
    dataframe=train,
    x_col='Filepath',
    y_col='Label',
    target_size=IMG_SIZE,
    class_mode='categorical', # One-hot encoding
    batch_size=BATCH_SIZE,
    shuffle=True,
    seed=SEED
)

```

```

valid_gen = valid_datagen.flow_from_dataframe(
    dataframe=valid,
    x_col='Filepath',
    y_col='Label',
    target_size=IMG_SIZE,
    class_mode='categorical',
    batch_size=BATCH_SIZE,
    shuffle=False,
    seed=SEED
)

test_gen = test_datagen.flow_from_dataframe(
    dataframe=test,
    x_col='Filepath',
    y_col='Label',
    target_size=IMG_SIZE,
    class_mode='categorical',
    batch_size=BATCH_SIZE,
    shuffle=False
)

```

```

Training set: 10800 images
Validation set: 3600 images
Test set: 3600 images
Found 10800 validated image filenames belonging to 3 classes.
Found 3600 validated image filenames belonging to 3 classes.
Found 3600 validated image filenames belonging to 3 classes.

```


Appendix B - Fine-tune DenseNet121 Architecture

Define the DenseNet121 Model Architecture

```
from tensorflow.keras.applications import DenseNet121
from tensorflow.keras import layers, models
from tensorflow.keras.optimizers import Adam
from tensorflow.keras.regularizers import l2
from tensorflow.keras.callbacks import ReduceLROnPlateau, EarlyStopping

# Load pre-trained DenseNet121
base_model = DenseNet121(weights="imagenet", include_top=False, input_shape=(224, 224, 3))

# Unfreeze only the last 50 layers for fine-tuning
for layer in base_model.layers[:-50]:
    layer.trainable = False

# Build the model
model = models.Sequential([
    base_model,
    layers.GlobalAveragePooling2D(),
    layers.Dense(256, activation="relu", kernel_regularizer=l2(0.0001)), # L2 Regularization (weaker)
    layers.Dropout(0.5), # Dropout for regularization
    layers.Dense(3, activation="softmax") # Updated for 3 classes
])

# Compile the model
model.compile(optimizer=Adam(learning_rate=1e-5), # Lower Learning rate for fine-tuning
              loss="categorical_crossentropy",
              metrics=["accuracy"])
```

```
# Learning rate scheduler & early stopping
from tensorflow.keras.callbacks import ReduceLROnPlateau, EarlyStopping

lr_scheduler = ReduceLROnPlateau(
    monitor='val_loss',
    factor=0.5,
    patience=3,
    min_lr=1e-7,
    verbose=1
)

early_stopping = EarlyStopping(
    monitor='val_loss',
    patience=5,
    restore_best_weights=True,
    verbose=1
)

# Model summary
model.summary()

Model: "sequential_3"
```

Model: "sequential_3"

Layer (type)	Output Shape	Param #
densenet121 (Functional)	(None, 7, 7, 1024)	7,037,504
global_average_pooling2d_3 (GlobalAveragePooling2D)	(None, 1024)	0
dense_6 (Dense)	(None, 256)	262,400
dropout_3 (Dropout)	(None, 256)	0
dense_7 (Dense)	(None, 3)	771

Total params: 7,300,675 (27.85 MB)
Trainable params: 431,299 (1.65 MB)
Non-trainable params: 6,869,376 (26.20 MB)

Appendix C - Structure Invariant Attack Algorithm

Algorithm 1: Structure Invariant Attack

Input: Classifier $f(\cdot)$ with the loss function J ; The benign sample \mathbf{x} with ground-truth label y ; The maximum perturbation ϵ , number of iterations T and decay factor μ ; Splitting number s ; Number of transformed images N

Output: An adversarial example.

```

1  $\alpha = \epsilon/T, \mathbf{g}_0 = 0, \mathbf{x}_0^{adv} = \mathbf{x}$ 
2 for  $t = 0 \rightarrow T - 1$  do
3   Constructing a set  $\mathcal{X}$  of  $N$  transformed images
   using SIT
4   Calculating the average gradient on  $\mathcal{X}$ :
      
$$\bar{\mathbf{g}}_{t+1} = \frac{1}{N} \sum_{\mathbf{x}_i \in \mathcal{X}} \nabla_{\mathbf{x}} J(\mathbf{x}_i, y) \quad (2)$$

5   Updating the momentum:
      
$$\mathbf{g}_{t+1} = \mu \mathbf{g}_t + \frac{\bar{\mathbf{g}}_{t+1}}{\|\bar{\mathbf{g}}_{t+1}\|_1} \quad (3)$$

6   Updating the adversarial example:
      
$$\mathbf{x}_{t+1}^{adv} = \text{Clip}(\mathbf{x}_t^{adv} + \alpha \cdot \text{sign}(\mathbf{g}_{t+1}), 0, 1) \quad (4)$$

7 return  $\mathbf{x}_T^{adv}$ 
```

Appendix D - SIA Attack hybrid defense measure algorithm

Defense Mechanism: Hybrid Approach (JPEG Compression + TVM)

Given an adversarial image I , a two-step defense is applied to restore and purify the image.

Step 1: JPEG Compression

The adversarial image is compressed using JPEG encoding and decoding to eliminate high-frequency perturbations:

$$I^{\text{JPEG}} = \text{Decode}(\text{Encode}(I, q))$$

where:

- I : original adversarial image
- q : JPEG quality parameter (e.g., 50)

Step 2: Total Variation Minimization (TVM)

Total variation minimization is then applied to further suppress residual noise:

$$I^{\text{TVM}} = \arg \min_{I'} \{ \|I^{\text{JPEG}} - I'\|^2 + \lambda \cdot \text{TV}(I') \}$$

The total variation term $\text{TV}(I')$ is defined as:

$$\text{TV}(I') = \sum_{i,j} \sqrt{(I'_{i+1,j} - I'_{i,j})^2 + (I'_{i,j+1} - I'_{i,j})^2}$$

where:

- λ : regularization parameter controlling the smoothness (e.g., 0.1)
- I' : the denoised image candidate

Final Defended Image

The final defended image after applying both steps is:

$$I^{\text{defended}} = \text{TVM}(\text{JPEG}(I)) = I^{\text{TVM}}$$

Appendix E - SIA attack detection algorithm

Equation for Attack Detection:

$$\text{SSIM} = \text{SSIM}(O_{\text{gray}}, A_{\text{gray}})$$

$$\text{Confidence Drop} = C_O - C_A$$

$$\text{Block Difference} = \frac{1}{N} \sum_{i=1}^N |O_{\text{gray}}(i) - A_{\text{gray}}(i)|$$

$$\text{Detection} = \begin{cases} \text{Attack Detected,} & \text{if } (\text{SSIM} < 0.85) \vee (\text{Confidence Drop} > 15\%) \vee (\text{Block Difference} > 20) \vee (\hat{Y}_O \neq \hat{Y}_A) \\ \text{No Attack,} & \text{otherwise} \end{cases}$$

Where:

- O and A are the **original** and **adversarial** images.
- O_{gray} and A_{gray} are their grayscale versions.
- $\text{SSIM}(O, A)$ computes the **structural similarity index**.
- C_O and C_A are the **classification confidence scores** before and after the attack.
- \hat{Y}_O and \hat{Y}_A are the **predicted labels** before and after the attack.
- N is the number of pixels in the image.