

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

24-25J-079

Project Proposal Report

Sandamal P.G.E.J

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

Department of Computer Systems and Engineering

Faculty of Computing

Sri Lanka Institute of Information Technology

Date of submission

August 2024

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

24-25J-079

Project Proposal Report

Sandamal P.G.E.J

IT21166860

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

Department of Computer Systems and Engineering

Faculty of Computing

Sri Lanka Institute of Information Technology

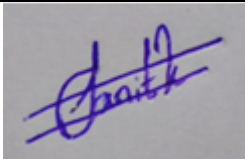
Date of submission

August 2024

DECLARATION

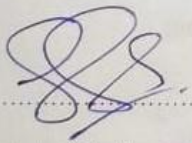
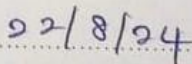
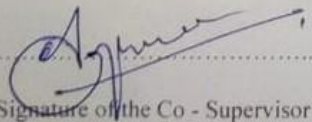
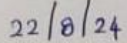
We declare that this is our own work, and this proposal does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or Institute of higher learning and to the best of our knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Group Member

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

The supervisor/s should certify the proposal report with the following declaration.

The above candidates are carrying out research for the undergraduate Dissertation under my supervision.

	
Signature of the Supervisor	Date
	
Signature of the Co - Supervisor	Date

ABSTRACT

The rapid advancement of deep learning, particularly Convolutional Neural Networks (CNNs), has revolutionized medical imaging, significantly improving the accuracy of diagnoses through chest X-rays. However, as reliance on these models grows, so does their vulnerability to adversarial attacks. These attacks subtly alter input images, leading CNNs to make incorrect diagnoses—potentially with serious consequences in medical settings. This research proposal focuses on the robustness of CNN models used in chest X-ray classification against the Split Image Adversarial (SIA) attack. The SIA attack involves dividing an image into segments, with each segment being subtly altered in a way that is nearly imperceptible to the human eye but can mislead the CNN, resulting in incorrect outputs. The study will begin by implementing this attack on a standard CNN model to assess its impact on diagnostic accuracy.

Following this, the research will focus on developing and testing defense strategies to enhance CNN resilience. Two key approaches will be explored: data augmentation, which expands the training dataset with variations to help the model resist attacks, and advanced adversarial training, where the model is trained to recognize and counteract adversarial inputs. These defenses will be rigorously evaluated to determine their effectiveness in improving CNN robustness, ensuring accurate diagnostics even under sophisticated attacks like SIA. The goal is to provide practical insights into safeguarding medical imaging systems against adversarial threats, contributing to more secure and trustworthy AI-driven healthcare solutions.

Keywords : *CNN robustness, Defense Strategies, Image Segmentation, Splitting Image Adversarial (SIA) attack.*

TABLE OF CONTENTS

DECLARATION	3
ABSTRACT.....	4
TABLE OF CONTENTS.....	5
TABLE OF FIGURES	7
TABLE OF TABLES	7
1. INTRODUCTION	8
1.1. Overview	8
1.2. Problem Statement	8
1.3. Research Questions	8
1.4. Significance of the Study	9
2. BACKGROUND AND LITERATURE SURVEY	10
2.1. CNNs in Medical Imaging	10
2.2. Adversarial Attacks on CNNs.....	10
2.3. Existing Defense Strategies	10
2.4. Gaps in Current Research	11
3. RESEARCH GAP.....	12
3.1. Identification of Research Gap	12
3.2. Justification for the Study	13
4. RESEARCH PROBLEM.....	14
4.1. Statement of the Research Problem	14
4.2. Scope of the Research.....	15
5. OBJECTIVES	16
5.1. Main Objectives	16
5.2. Specific Objectives	16
5.2.1. Implement SIA Attack	16
5.2.2. Fine-Tune CNN Model	16
5.2.3. Test Defense Strategies	17
5.2.4. Evaluate Performance Metrics	17
5.2.5. Validate Defense Effectiveness	17
6. METHODOLOGY	18
6.1. Overall System Diagram.....	18
6.3. System Diagram for Component	19
6.2. Technologies	19
6.3. Work Breakdown Structure	21
7. PROJECT REQUIREMENTS.....	23

7.1. Hardware Requirements.....	23
7.2. Software Requirements.....	23
7.3. Data Requirements.....	24
8.GANTT CHART	25
9. BUDGET AND BUDGET JUSTIFICATION	27
9.2 Hardware Costs	27
9.3 Cloud Computing Resources	28
9.4 Software Costs	28
10. REFERENCES	29

TABLE OF FIGURES

Figure 4.1 : Raw Image vs SIA attack affected image.....	
Figure 6.1 Overall System Diagram.....	
Figure 6.2 SIA attack Component-Specific System Diagram.....	
Figure 6.3 Work Breakdown Diagram.....	
Figure 8.1 Gantt Chart.....	

TABLE OF TABLES

Table 3.1: Identified Research Gaps in CNN Robustness.....	
Table 2.1 : Budget Allocation Table.....	

1. INTRODUCTION

1.1. Overview

The rapid integration of Artificial Intelligence (AI) into healthcare has transformed medical diagnostics, particularly with the use of Convolutional Neural Networks (CNNs) in medical imaging. CNNs have shown remarkable accuracy in interpreting medical images, such as chest X-rays, aiding in the early detection and diagnosis of diseases. However, the robustness of these models is increasingly questioned, particularly in the face of adversarial attacks. These attacks, which subtly alter the input images, can lead to significant misdiagnoses, posing a critical challenge in medical diagnostics. This research focuses on assessing the robustness of CNN models in the medical imaging domain, specifically against the Split Image Adversarial (SIA) attack and exploring effective defense strategies to enhance their reliability.

1.2. Problem Statement

Despite the advancements in CNNs for medical imaging, these models remain vulnerable to adversarial attacks that can compromise their diagnostic accuracy. The Split Image Adversarial (SIA) attack, which involves segmenting and subtly altering parts of the image, is a particularly concerning threat. This attack can lead to incorrect diagnoses, which may have severe consequences in clinical settings. The problem this research addresses is the need to assess and enhance the robustness of CNN models against such adversarial threats, ensuring that medical diagnostics remain reliable and secure.

1.3. Research Questions

This study is guided by the following key research questions:

1. How effective are existing defense strategies in enhancing the robustness of CNN models used for chest X-ray classification against SIA attacks?
2. What specific approaches can be developed and tested to further strengthen CNN models against these adversarial threats?
3. How can the effectiveness of these defense strategies be validated to ensure secure and reliable medical diagnostics?

1.4. Significance of the Study

The significance of this study lies in its potential to contribute to safer and more reliable AI-driven medical diagnostics. By focusing on the Split Image Adversarial (SIA) attack, this research addresses a critical gap in the existing literature on CNN robustness in medical imaging. The outcomes of this study are expected to offer practical solutions for enhancing the security and accuracy of CNN models, ultimately benefiting both healthcare providers and patients by reducing the risk of misdiagnoses caused by adversarial attacks.

2. BACKGROUND AND LITERATURE SURVEY

2.1. CNNs in Medical Imaging

Convolutional Neural Networks (CNNs) have revolutionized the field of medical imaging by providing advanced capabilities in image analysis and classification. CNNs leverage multiple layers of convolutions, pooling, and activation functions to automatically extract and learn features from medical images, such as chest X-rays. This capability has led to significant improvements in diagnostic accuracy, enabling more precise identification of various conditions, including pneumonia and tuberculosis. Recent studies have demonstrated that CNNs can outperform traditional image analysis methods, providing higher sensitivity and specificity in medical image classification [1].

2.2. Adversarial Attacks on CNNs

Despite their effectiveness, CNNs are vulnerable to adversarial attacks, which involve subtly altering input images to mislead the model's predictions. These attacks can be particularly dangerous in medical imaging, where small perturbations can lead to incorrect diagnoses. The Split Image Adversarial (SIA) attack is a recent advancement in this area, where the image is divided into segments, and adversarial perturbations are applied to these segments individually. This type of attack can exploit the CNN's weaknesses, leading to significant errors in medical image classification [2], [3].

2.3. Existing Defense Strategies

To combat adversarial attacks, various defense strategies have been proposed. These include data augmentation, which involves expanding the training dataset with modified versions of images to improve model robustness, and adversarial training, where the model is trained on both clean and adversarial perturbed images to enhance its resistance to attacks. Other approaches include robust optimization techniques, and the use of defensive neural network architectures designed to detect and mitigate adversarial perturbations. While these methods show promise, their effectiveness can vary depending on the nature of the adversarial attack and the complexity of the CNN model [4], [5].

2.4. Gaps in Current Research

Despite advancements in defense strategies, there are still significant gaps in current research. Many existing methods do not adequately address specific adversarial techniques like the SIA attack, and there is limited research on how these defenses perform in practical medical imaging scenarios. Additionally, there is a need for comprehensive evaluations of defense strategies that consider both the performance impact on CNN models and the practical implications for medical diagnostics. Addressing these gaps is crucial for developing more robust and reliable AI systems for healthcare [6], [7].

3. RESEARCH GAP

3.1. Identification of Research Gap

The study of Convolutional Neural Networks (CNNs) in medical imaging has made significant strides; however, several critical research gaps remain, particularly in the context of adversarial attacks such as the Split Image Adversarial (SIA) attack. The key gaps identified are:

1. **In-depth Robustness Assessment of CNN Models Against SIA Attacks in Medical Imaging Systems:** Previous research has addressed adversarial attacks in general, but there is a lack of detailed robustness assessments specific to SIA attacks in medical imaging systems. This gap highlights the need for a focused evaluation of how CNN models respond to SIA attacks and the effectiveness of various defense mechanisms in this specific context.
2. **Focus on Understanding and Mitigating the Impact of SIA Attacks on Chest X-ray Classification Models:** While some studies have explored adversarial attacks on medical images, few have concentrated on the SIA attack specifically and its impact on chest X-ray classification models. This area remains under-explored, making it crucial to investigate how SIA attacks affect diagnostic accuracy and to develop tailored defense strategies.
3. **Evaluation on Medical Image Datasets, Providing Insights Specific to Medical Image Classification Systems:** Research has been conducted on adversarial attacks and defenses, but detailed evaluations on specific medical image datasets, such as chest X-rays, are limited. This gap underscores the need for research that provides insights specific to medical image classification systems and assesses the effectiveness of defenses in this domain.
4. **Basic Data Augmentation Techniques Used in Conjunction with Adversarial Defenses:** While various data augmentation techniques have been proposed, their effectiveness when used alongside adversarial defenses remains inadequately studied. This gap presents an opportunity to explore basic data augmentation techniques and their role in enhancing the robustness of CNN models against adversarial attacks.

5. **Comprehensive Performance Metrics Under SIA Attack:** There is a lack of comprehensive performance metrics that specifically address the effectiveness of defenses against attacks such as SIA (Splitting Image Adversarial) attacks. This gap highlights the need for developing and applying detailed performance metrics to evaluate the resilience of CNN models under such attacks.





















	Research 1	Research 2	Research 3	Our Research
In-depth robustness assessment of CNN models against SIA attacks in the context of medical imaging systems.				
Focus on understanding and mitigating the impact of SIA attacks on chest X-ray classification models.				
Evaluation on medical image datasets, providing insights specific to medical image classification systems.				
Basic data augmentation techniques used in conjunction with adversarial defences				
Comprehensive performance metrics under SIA attack				

Table 3.1: Identified Research Gaps in CNN Robustness

3.2. Justification for the Study

Addressing the identified research gaps is crucial for advancing the field of medical imaging and improving the reliability of CNN models in clinical settings. By conducting an in-depth assessment of CNN robustness against SIA attacks, this study aims to fill a significant void in current research. Understanding and mitigating the impact of these attacks on chest X-ray classification models will contribute to more accurate and reliable diagnostic tools. Additionally, evaluating basic data augmentation techniques and developing comprehensive performance metrics will provide valuable insights into enhancing the robustness of CNN models. This research not only aims to advance theoretical knowledge but also to offer practical solutions that can be directly applied to improve medical diagnostic systems.

4. RESEARCH PROBLEM

4.1. Statement of the Research Problem

The integration of Convolutional Neural Networks (CNNs) into medical imaging has significantly advanced diagnostic capabilities. However, these models are not without their vulnerabilities, particularly when faced with adversarial attacks. One such threat is the Split Image Adversarial (SIA) attack, which targets CNNs by segmenting medical images and introducing subtle perturbations that can compromise diagnostic accuracy. Despite the progress made in developing CNN-based diagnostic tools, there is a critical gap in understanding how these models withstand specific adversarial attacks like SIA, particularly in the context of chest X-ray classification. This vulnerability poses a serious risk to the reliability of AI-driven medical diagnostics, potentially leading to misdiagnoses and affecting patient care.

Current research has highlighted the general impact of adversarial attacks on CNNs, but there is a notable lack of comprehensive studies focusing on the SIA attack and its effects on medical imaging systems. Additionally, while various defense strategies have been proposed, their effectiveness in countering specific attacks like SIA remains underexplored. Addressing this research problem involves not only assessing the robustness of CNN models against SIA attacks but also developing and validating effective defense strategies to enhance model reliability and ensure secure medical diagnostics.



Figure 4.1 : Raw Image vs SIA attack affected image

4.2. Scope of the Research

This research focuses on the application of Convolutional Neural Networks (CNNs) in medical imaging, with a specific emphasis on chest X-ray classification. The study will primarily address the vulnerabilities of CNN models to the Split Image Adversarial (SIA) attack and will evaluate various defense strategies to mitigate these vulnerabilities. The scope includes:

- **Assessment of CNN Models:** The research will involve a detailed analysis of CNN models used for chest X-ray classification, assessing their susceptibility to SIA attacks.
- **Development of Defense Mechanisms:** The study will explore and implement defense strategies, including data augmentation and adversarial training, to enhance the robustness of CNN models.
- **Evaluation and Validation:** The effectiveness of these defense mechanisms will be evaluated through performance metrics and real-world testing, providing insights into their impact on diagnostic accuracy.
- **Contextual Application:** While the research will focus on chest X-ray images, the findings are expected to be applicable to other areas of medical imaging where CNNs are employed.

The research is designed to contribute to the development of more robust and reliable AI systems in healthcare, addressing both theoretical and practical aspects of adversarial resilience in medical imaging.

5. OBJECTIVES

5.1. Main Objectives

The primary aim of this research is to enhance the robustness of Convolutional Neural Networks (CNNs) used in medical imaging against adversarial threats, specifically the Split Image Adversarial (SIA) attack. This research seeks to achieve a comprehensive understanding of how CNN models can be made more resilient to such attacks, with a focus on chest X-ray classification systems. The main objectives include:

1. **Implementing and Evaluating Defenses:** Develop and apply various defense strategies to counteract the SIA attack. This involves testing and validating these defenses to assess their effectiveness in improving the robustness of CNN models.
2. **Assessing Model Robustness:** Conduct a thorough assessment of CNN models under attack conditions to identify vulnerabilities and evaluate the impact of the implemented defenses on diagnostic accuracy and model performance.

By addressing these objectives, the research aims to bridge the gap between theoretical advancements and practical applications, ultimately contributing to the development of more secure and reliable medical imaging systems.

5.2. Specific Objectives

To achieve the main objectives, the research will focus on the following specific goals:

5.2.1. Implement SIA Attack

The first step involves implementing the Split Image Adversarial (SIA) attack on CNN models used for chest X-ray classification. This task includes developing a robust methodology for applying the SIA attack, where medical images are split into segments, and perturbations are introduced to test the models' vulnerability. The implementation will provide a baseline for evaluating how different CNN architectures respond to adversarial manipulations.

5.2.2. Fine-Tune CNN Model

After establishing a baseline with the SIA attack, the next objective is to fine-tune the CNN models to improve their robustness. This process involves adjusting model parameters and training techniques to enhance their resistance to adversarial attacks. The fine-tuning process will be guided by insights gained from the initial implementation of the SIA attack.

5.2.3. Test Defense Strategies

With fine-tuned models in place, the research will then focus on testing various defense strategies designed to mitigate the effects of the SIA attack. These strategies may include data augmentation, adversarial training, and other robust optimization techniques. The effectiveness of each defense strategy will be evaluated to determine its impact on model performance and resilience.

5.2.4. Evaluate Performance Metrics

To gauge the success of the defense strategies and fine-tuned models, a comprehensive evaluation of performance metrics will be conducted. This includes assessing metrics such as accuracy, precision, recall, and robustness against adversarial examples. The evaluation will help determine how well the models and defenses perform in real-world scenarios and provide insights into their practical applicability.

5.2.5. Validate Defense Effectiveness

Finally, the research will validate the effectiveness of the implemented defense strategies through rigorous testing and analysis. This involves comparing the performance of defended models against those without defenses and ensuring that the proposed strategies effectively enhance the robustness of CNN models. Validation will include both quantitative assessments and qualitative evaluations to ensure that the defenses provide meaningful improvements in model security and reliability.

6. METHODOLOGY

6.1. Overall System Diagram

The overall system diagram outlines the high-level architecture and workflow of the research project. This diagram illustrates how various components interact, from data acquisition to model evaluation. The diagram includes:

1. **Data Collection:** Chest X-ray images are collected and pre-processed for input into the Convolutional Neural Network (CNN) models.
2. **Attack Implementation:** The Split Image Adversarial (SIA) attack is applied to the pre-processed images to create adversarial examples.
3. **Model Training and Fine-Tuning:** CNN models are trained and fine-tuned on the original and adversarial datasets.
4. **Defense Strategies Application:** Various defense mechanisms, such as data augmentation and adversarial training, are applied to enhance model robustness.
5. **Evaluation:** The performance of the models, both with and without defenses, is evaluated using metrics to assess accuracy, robustness, and resilience to adversarial attacks.

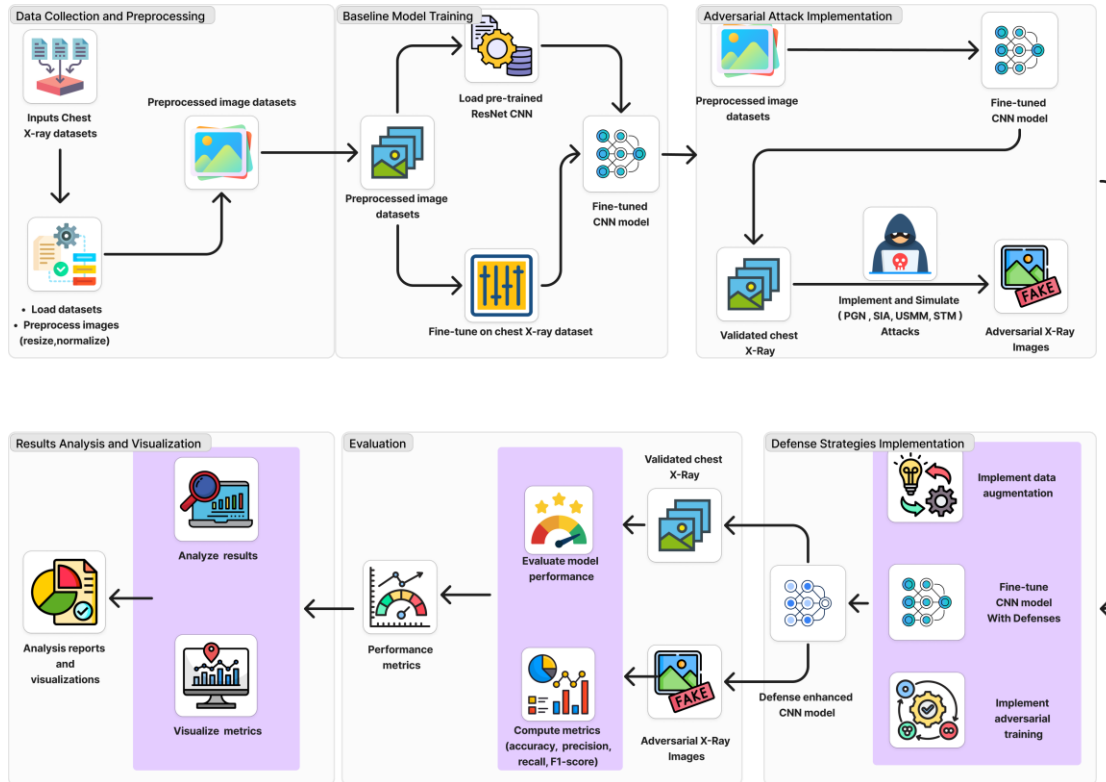


Figure 6.1 Overall System Diagram

6.3. System Diagram for Component

This system diagram specific to my research component focuses on the implementation of the Split Image Adversarial (SIA) attack and the corresponding defense strategies. It includes:

1. **SIA Attack Implementation:** Detailed steps of segmenting the medical images and introducing perturbations to test the CNN models' vulnerabilities.
2. **Model Fine-Tuning:** Process of adjusting CNN models to improve their robustness against SIA attacks.
3. **Defense Strategies Testing:** Application of various defense mechanisms and their integration into the model pipeline.
4. **Performance Evaluation:** Analysis of the model's performance using pre-defined metrics to validate the effectiveness of the defenses.

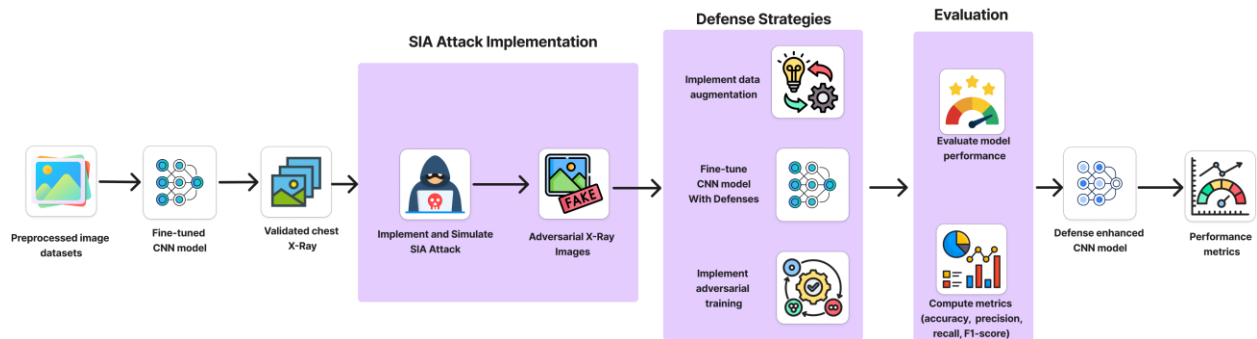


Figure 6.2 SIA attack Component-Specific System Diagram

6.2. Technologies

This section describes the technologies and tools used in the research, providing a comprehensive overview of the resources and frameworks that support the methodology.

Deep Learning Frameworks:

- **TensorFlow:** An open-source library for machine learning and deep learning that provides a flexible and comprehensive ecosystem for building and training CNN models.
- **PyTorch:** Another open-source deep learning framework that offers dynamic computation graphs and a user-friendly interface for model development and training.

Programming Languages:

- **Python:** The primary programming language used for implementing the models, attacks, and defenses. Python's extensive libraries and frameworks make it ideal for deep learning and data analysis.

Model Architecture:

- **CNN (Convolutional Neural Network):** The architecture used for medical image classification. CNNs are particularly effective for processing and analyzing visual data due to their ability to capture spatial hierarchies.

Data Management Tools:

- **Pandas:** A powerful data manipulation and analysis library that facilitates handling and preprocessing of large datasets.
- **NumPy:** A fundamental package for scientific computing with Python, providing support for large, multi-dimensional arrays and matrices.

Visualization Tools:

- **Matplotlib:** A plotting library used for creating static, animated, and interactive visualizations in Python.
- **Seaborn:** A statistical data visualization library based on Matplotlib that provides a high-level interface for drawing attractive and informative graphics.

Evaluation Metrics:

- **Scikit-learn:** A machine learning library that provides tools for model evaluation, including metrics such as accuracy, precision, recall, and F1-score.

IDE:

- **Google Colab:** An online integrated development environment (IDE) that supports Python and provides free access to GPU resources, making it suitable for training and testing deep learning models.

Hardware & Compute Resources:

- **Google Cloud:** Cloud computing resources used for scalable and efficient processing, including virtual machines with GPU support for training CNN models.

6.3. Work Breakdown Structure

The Work Breakdown Structure (WBS) organizes the project into manageable sections, each with defined tasks and deliverables. It includes:

1. **Project Planning:**
 - Define project scope and objectives
 - Develop detailed project schedule and milestones
2. **Data Preparation:**
 - Collect and preprocess chest X-ray images
 - Split data into training, validation, and test sets
3. **Attack Implementation:**
 - Develop and apply the SIA attack on medical images
 - Generate adversarial examples for model testing
4. **Model Training and Fine-Tuning:**
 - Train initial CNN models on clean images
 - Fine-tune models using adversarial examples
5. **Defense Strategies Development:**
 - Implement defense mechanisms such as data augmentation and adversarial training
 - Integrate defenses into the model pipeline
6. **Performance Evaluation:**
 - Assess model performance using evaluation metrics
 - Analyze the impact of defense strategies on model robustness
7. **Validation and Reporting:**
 - Validate the effectiveness of defense strategies
 - Prepare final report and documentation

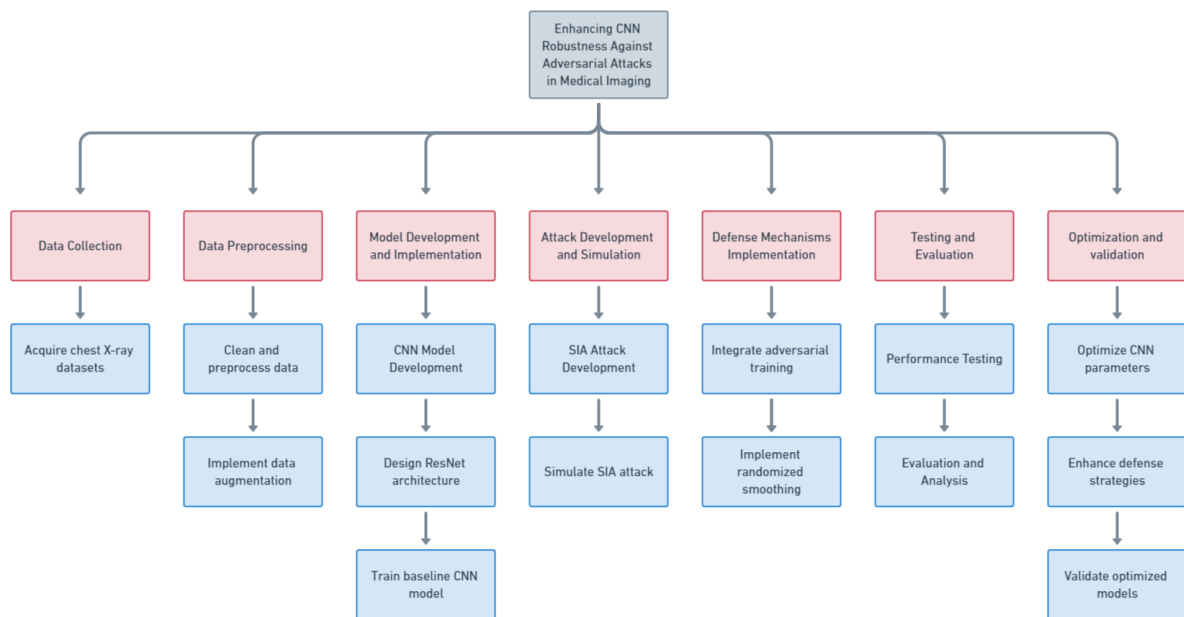


Figure 6.3 Work Breakdown Diagram

7. PROJECT REQUIREMENTS

7.1. Hardware Requirements

The research project demands robust hardware to efficiently handle the computationally intensive tasks associated with deep learning, particularly in the context of CNN training and testing under adversarial conditions.

- **High-Performance GPUs:** The use of GPUs such as NVIDIA Tesla or RTX series is crucial for accelerating the training of CNN models, especially when working with large datasets like chest X-ray images.
- **RAM:** A minimum of 16GB of RAM is necessary to manage large datasets and complex model architectures during the training process.
- **Storage:** A Solid State Drive (SSD) with at least 500GB of storage capacity is recommended to store datasets, models, and other essential project files, ensuring quick data access and processing.
- **Network:** High-speed internet connectivity is required for efficient data transfer, particularly when using cloud resources or accessing remote datasets.

7.2. Software Requirements

The software requirements include development environments, deep learning frameworks, and various libraries necessary for implementing, training, testing, and evaluating the CNN models and adversarial attacks.

- **Development Environment:**
 - **Python:** The primary programming language for implementing the models and attacks due to its extensive libraries and frameworks.
 - **Google Colab:** An online IDE that supports Python, offering free access to GPU resources for deep learning tasks.
- **Deep Learning Frameworks:**
 - **TensorFlow:** A widely-used open-source framework that provides extensive tools for building, training, and deploying deep learning models.
 - **PyTorch:** Another powerful deep learning framework known for its dynamic computation graphs and ease of use in research settings.

- **Libraries and Tools:**
 - **Adversarial Attack Libraries:** Libraries specifically designed to facilitate the implementation of adversarial attacks on deep learning models.
 - **Data Management:**
 - **Pandas:** For data manipulation and preprocessing, essential for handling large datasets effectively.
 - **NumPy:** For numerical computations, enabling efficient operations on large arrays and matrices.
 - **Visualization:**
 - **Matplotlib:** For creating static, animated, and interactive visualizations to analyze model performance and data characteristics.
 - **Seaborn:** For producing statistical data visualizations that provide deeper insights into the dataset and model performance.
 - **Scikit-learn:** For evaluating model performance using a variety of metrics, and for additional machine learning tasks such as classification and regression.

7.3. Data Requirements

The success of this research heavily depends on the availability and quality of the data used for training, testing, and validating the CNN models.

- **Chest X-ray Image Dataset:** A large and diverse dataset of chest X-ray images is required to train the CNN models. The dataset must include normal and abnormal cases to ensure the model can generalize well.
- **Adversarial Data Generation:** Synthetic data generated through the application of the SIA attack, used to test the robustness of the CNN models and to evaluate the effectiveness of the defense strategies.
- **Data Preprocessing Tools:** Tools and scripts for preprocessing the raw data, such as resizing, normalization, and augmentation, to prepare it for input into the CNN models.

8. GANTT CHART

The Gantt chart below outlines the timeline for the research project, presenting the sequence of tasks, their duration, and key milestones. The project is divided into distinct phases that align with the methodology, each critical to the successful completion of the research.

Phase 1: Project Planning

- Define the project scope and objectives
- Develop a detailed project schedule and identify key milestones
- Set up the development environment (Google Colab, TensorFlow, PyTorch)

Phase 2: Data Preparation

- Collect and preprocess the chest X-ray dataset
- Split data into training, validation, and test sets
- Implement data augmentation techniques

Phase 3: Model Development

- Develop and train initial CNN models on clean images
- Fine-tune models to improve baseline performance
- Document model architecture and parameters

Phase 4: SIA Attack Simulation

- Implement the SIA attack on medical images
- Generate adversarial examples to test against the CNN models
- Analyze the impact of the attack on model performance

Phase 5: Defense Strategy Implementation

- Develop and integrate various defense mechanisms (e.g., data augmentation, adversarial training)
- Test defenses against the SIA attack
- Document the defense implementation process

Phase 6: Testing and Evaluation

- Evaluate model performance using predefined metrics (e.g., accuracy, precision, recall)
- Compare the model's robustness before and after applying defense strategies
- Compile test results for further analysis

Phase 7: Optimization and Refinement

- Optimize defense strategies based on initial evaluation results
- Refine the model and defense mechanisms to enhance robustness
- Conduct additional testing to validate improvements

Phase 8: Documentation

- Compile all research findings and results
- Prepare the final report and presentation
- Submit documentation and present findings

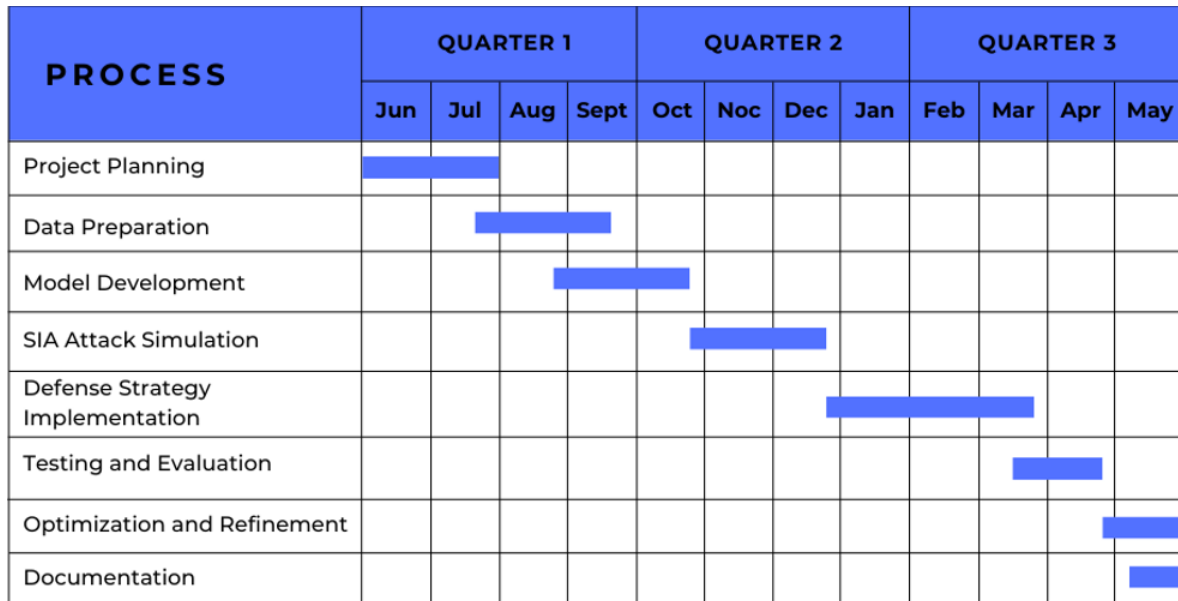


Figure 8.1 Gantt Chart

9. BUDGET AND BUDGET JUSTIFICATION

This section outlines the estimated costs associated with the research project, categorized into hardware, cloud computing resources, and software costs. Each budget item is carefully considered to ensure the successful execution of the project while optimizing resource allocation.

Description	Cost
Hardware Costs	
High-Performance GPUs	LKR 50,000 – LKR 1,00,000
RAM (Minimum 16GB)	LKR 6,000 – LKR 8,000
Storage (SSD with at least 500GB)	LKR 5,000 – LKR 7,000
Cloud Computing Resources	
Google Colab Pro subscription	Approximately LKR 1,000 per month
Software Costs Development Environment and Libraries	Free

Table 2.1 : Budget Allocation Table

9.2 Hardware Costs

1. High-Performance GPUs:

- **Cost Estimate:** LKR 50,000 – LKR 1,00,000
- **Justification:** The use of high-performance GPUs is essential for the efficient training of CNN models, especially when dealing with large datasets like chest X-ray images. GPUs significantly reduce training time, allowing for more iterative testing and refinement of models, which is crucial for the project's success.

2. RAM (Minimum 16GB):

- **Cost Estimate:** LKR 6,000 – LKR 8,000
- **Justification:** Adequate RAM is necessary to handle the memory-intensive tasks involved in training deep learning models. With at least 16GB of RAM, the system can process large batches of data without running into memory bottlenecks, ensuring smooth and efficient operation during model training and testing.

3. Storage (SSD with at least 500GB):

- **Cost Estimate:** LKR 5,000 – LKR 7,000

- **Justification:** A fast and reliable SSD with sufficient storage capacity is vital for storing large datasets, models, and other project files. The quick access speeds of SSDs will enhance data processing times and overall workflow efficiency, particularly when managing large amounts of image data.

9.3 Cloud Computing Resources

1. Google Colab Pro Subscription:

- **Cost Estimate:** Approximately LKR 1,000 per month
- **Justification:** Google Colab Pro provides access to enhanced computational resources, including faster GPUs and more extended runtime sessions. This subscription is essential for running deep learning experiments, especially those requiring prolonged training times, and for utilizing advanced features not available in the free tier.

9.4 Software Costs

1. Development Environment and Libraries:

- **Cost Estimate:** Free
- **Justification:** The primary software tools, including Python, TensorFlow, PyTorch, and various data management and visualization libraries, are open-source and freely available. This allows the project to leverage powerful tools without incurring additional software costs, ensuring cost-efficiency while maintaining high functionality.

10. REFERENCES

- [1] A. Esteva, B. Kuprel, R. Novoa, et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115-118, Jan. 2023.
- [2] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.
- [3] N. Papernot, P. McDaniel, and A. Sinha, "The limitations of deep learning in adversarial settings," *Proceedings of the IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 372-387, Mar. 2024.
- [4] X. Zhang, S. Wang, and Y. Chen, "Defending against adversarial attacks using robust optimization," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1600-1608, Jun. 2023.
- [5] C. Szegedy, W. Zaremba, I. Sutskever, et al., "Intriguing properties of neural networks," *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.
- [6] B. Biggio, I. Corona, D. Fumera, and F. Roli, "Evasion attacks against machine learning at test time," *Proceedings of the 29th International Conference on Machine Learning (ICML)*, vol. 28, pp. 180-187, Jul. 2024.
- [7] A. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.
- [8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 770-778. Available: <https://ieeexplore.ieee.org/document/7780459>
- [9] N. Papernot et al., "The limitations of deep learning in adversarial settings," in *Proceedings of the IEEE European Symposium on Security and Privacy (EuroS&P)*, 2019, pp. 372-387. Available: <https://ieeexplore.ieee.org/document/7514792>
- [10] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. Available: <https://arxiv.org/abs/1706.06083>
- [11] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. Available: <https://arxiv.org/abs/1611.01236>