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

Premakanthan N

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

Department of Computer Systems Engineering

Sri Lanka Institute of Information Technology Sri Lanka

August 2024

ASSESSING CNN ROBUSTNESS TO STYLE TRANSFER MANIPULATION (STM) ADVERSARIAL ATTACKS

Project Proposal Report

Premakanthan N

IT21197550

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

Department of Information Technology

Sri Lanka Institute of Information Technology Sri Lanka

August 2024

Declaration

We declare that this is our own work, and this proposal does not incorporate without acknowledgment 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 acknowledgment is made in the text.

Name	Student ID	Signature
Premakanthan N	IT21197550	

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

Supervisor: Dr. Harinda Fernando

.... 22/8/24

Signature Date

Co-Supervisor: Mr. Kavinga Abeywardena

.... 22/08/24

Signature Date

Abstract

Convolutional Neural Networks are among the most profoundly applied tools in medical imaging, particularly about tasks such as chest X-ray classification, where accuracy is required in diagnosing diseases. These models, however, could be vulnerable to a certain line of adversarial attacks that involve style transfer manipulation. STM attacks are a type of attack that subtly modifies image textures and colors to mislead CNNs into making incorrect classifications. This paper will focus on how CNN models handle STM attacks by testing the model on a chest X-ray dataset. We will tune the CNN models and check their performance by accuracy evaluation metrics, precision, recall, and the F1 score. Furthermore, we will test several defense strategies that help protect the models from STM attacks. These would include data augmentation, exposing the model to a wide variety of images during training, and another is adversarial training, which aims at training models that are intended to recognize and consequently resist such attacks. This will involve comparisons between the results for both clean and attacked datasets to infer the strongest defenses. One major contribution of this work is to improve the reliability and security of CNNs in medical imaging. That is, this research ensures that accurate diagnosis can still be made under any sophisticated adversarial attacks.

Keywords: *Convolutional Neural Networks (CNNs), Style Transfer Manipulation (STM), adversarial attacks, model robustness, defense mechanisms.*

Acknowledgment

I would like to express my deepest gratitude to my supervisor, Dr. Harinda Fernando, and my co-supervisor, Mr. Kavinga Abeywardena, for their invaluable guidance, support, and inspiration throughout this research journey. Their expertise and encouragement were instrumental in the successful completion of this project. I also extend my sincere thanks to the Department of Computer Systems Engineering at the Sri Lanka Institute of Information Technology and all the lecturers and staff for providing me with the opportunity and resources to conduct this research.

Table of Contents

Table of Figures

Table of Tables

List of Abbreviations

Abbreviation	**Description**
CNN	Convolutional Neural Networks
STM	Style Transfer Manipulation
CVPR	Computer Vision and Pattern Recognition
STR	Scene Text Recognition
IDE	Integrated Development Environment

1. INTRODUCTION

Medical imaging systems have been the mainstay of healthcare in the modern era for diagnosing several diseases. These are powered by advanced machine and deep learning models, especially Convolutional Neural Networks to help rightly identify conditions from medical images like X-rays, MRIs, and CT scans. For instance, CNN models like ResNet are normally used to read chest X-rays for diseases like pneumonia, lung cancer, and tuberculosis.[5]

However, the more these models become part of healthcare, the greater they also become as potential targets of adversarial attacks. An attacker can manipulate these models to change their predictions, which can be potentially disastrous. In adversarial attacks, only small, often imperceptible, changes in the input images are required, which can cause the model to misclassify them. This can be either positive, miss labeling a diseased patient as healthy, or negative, labeling a healthy patient with a disease. Both changes are risky and predispose to a potential danger to the patients through wrong treatment decisions and delayed proper care.

One of the particularly insidious types of adversarial attacks is called a Style Transfer Manipulation (STM) attack. STM attacks work by subtly changing the style elements of an image, such as its texture or color or other aesthetic features, without touching its overall structure or content. Quite often, the changes will be so subtle that they are imperceptible to the human eye but may result in deep learning models' failure to make predictions. For instance, as illustrated in the next section, when applied to medical imaging, an STM attack would create slight distortion in a chest X-ray image, leading the model to misinterpret it and miss a disease or diagnose a disease when there isn't one.

Given the critical nature of the roles that the models play in healthcare, their robustness against such attacks is of importance. The presented work aims to determine how robust CNN models are in classifying X-ray image classification tasks against STM attacks and investigating effective defense approaches to protecting these systems from adversarial manipulations.

2. BACKGROUND AND LITERATURE SURVEY

2.1. Overview

The project consists of several critical components that ensure the robustness of Convolutional Neural Networks (CNNs) used in medical imaging systems. The Data Input and Preprocessing component handles the ingestion and preparation of medical images, including resizing, normalization, and data augmentation. The CNN Model Architecture component defines and implements CNN models, constructing layers for analyzing medical images. The Advanced Attack Simulation component tests the robustness of these CNN models by generating adversarial examples to identify potential vulnerabilities. The Defense Mechanisms component implements strategies to protect the CNNs from such threats, such as adversarial training and defensive distillation.[6]

The Model Evaluation and Validation component assesses the models using metrics such as accuracy, precision, recall, and robustness scores, providing critical insights into their performance under normal and adversarial conditions. The Visualization and Reporting component generates charts, graphs, and reports to aid in interpreting and communicating findings.

The System Integration and Deployment component ensures the robust CNN models can be seamlessly integrated into existing clinical systems, enabling their deployment in real-world medical environments. The User Interface and Interaction component provides a user-friendly interface for medical professionals and researchers to interact with the system, upload images, select models, and view results without needing deep technical expertise.

In summary, the project aims to enhance the safety, reliability, and performance of CNNs in medical imaging systems, making them more resilient to adversarial threats and more effective in clinical applications.

2.2. Literature Survey

A comprehensive literature review was conducted to explore existing research on adversarial attacks and style transfer techniques, particularly in the context of machine learning models and medical imaging.

In 2016, Gatys, Ecker, and Bethge presented a groundbreaking work at the IEEE Conference on Computer CVPR on image style transfer using CNN. They introduced a Neural Algorithm of Artistic Style, which combines the content and style of images to create new images. Although primarily focused on aesthetic applications, the techniques are relevant to the development of STM attacks, as they allow for altering image styles without affecting the underlying content.[12]

The study by Zhilu Zhang, Xi Yang, and Kaizhu Huang in 2021 introduces an innovative approach to generate adversarial examples for sequential learning models like Scene Text Recognition (STR). They use a Style Transfer Mechanism to alter the style of an image without changing its perceptual content, misleading the model. This research shows that STR networks, including TRBA and CRNN, are susceptible to style-based adversarial attacks, significantly affecting their accuracy. However, the study does not extend its exploration to medical imaging contexts, suggesting a potential area for further research in applying these techniques to CNN models in healthcare.[1]

Rui Huang's 2022 paper explores the use of harmonized regional style transfer in medical images to enhance visual quality while maintaining diagnostic accuracy. The study suggests that these methodologies could be used to explore the resilience of medical imaging systems to style-based adversarial attacks. However, it does not address the adversarial potential of style transfer, suggesting a need for further research in developing STM techniques for adversarial attacks in medical imaging.[4]

The review paper by Li et al. in 2022 in the International Journal of Medical Informatics provides an overview of security threats faced by image classification systems in medical imaging. It emphasizes the importance of securing these systems against adversarial attacks and discusses existing defense mechanisms. However, the review does not address STM as a specific attack method, highlighting the need for targeted defense strategies.[13]

In summary, while the existing literature offers valuable insights into adversarial attacks and style transfer, there remains a significant gap in exploring these techniques within the context of medical imaging. Your research intends to bridge this gap by applying STM attacks to CNN models used in healthcare and developing robust defense mechanisms to ensure the reliability and security of these critical systems.

2.3. Research Gap

This research paper reports the vulnerabilities in Convolutional Neural Networks for medical imaging, specifically due to Style Transfer manipulation attacks. Most works in this area, while showing the potential vulnerabilities of STM at high levels, are conducted so that the implications for medical applications are not fully investigated. This work tests with due diligence the perturbations on CNN models used in the classification of chest X-rays. The work further investigates and contrasts different defense strategies to eventually find out which defense works better in guarding CNNs against such devastating attacks. With this background, the current work addresses the above-mentioned critical gaps and is going to better the CNN-based medical imaging systems with improvements over current research.

Research Paper	Adversarial Attacks	Style Transfer	Medical Images	Defense Strategies	Evaluation Metrics
Research 1 [1]	✓	✓	X	X	✓
Research 2 [11]	X	✓	X	X	X
Research 3 [3]	✓	✓	X	X	✓
Research 4 [4]	X	✓	✓	X	✓
Research 5 [7]	✓	X	✓	X	✓
Our Research	✓	✓	✓	✓	✓

Table 1: Research Gap

Adversarial Attacks: Whether the research addresses adversarial attacks specifically.

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

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

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

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

2.4. Problem Statement

Medical imaging systems, powered by Convolutional Neural Networks, apply state-of-the-art technology and are changing the face of health care by providing very accurate tools for the diagnosis from images like X-rays, MRIs, and CT scans. Such CNN models have become intrinsic to tasks, such as detecting anomalies in chest X-rays, which is one of the most important ways to diagnose ailments like pneumonia or lung cancer.

The issue with these CNN models is that they are extremely vulnerable, allowing themselves to be easily fooled by adversarial attacks. Among them, one of the most worrying is the STM attack. In an STM attack, the "style" of an image that is, its texture, color, or any other visual characteristics is subtly changed, while the principal content, like shape or structure, remains the same. These are often imperceptible changes to human eyes, yet a CNN model may make incorrect decisions, probably leading to misdiagnosis in clinical settings.

Specifically, STM attacks against CNN models used for medical imaging have not been investigated at all. That is a research gap, leaving our current medical imaging systems perhaps more vulnerable than we think. Moreover, although some defenses against these adversarial attacks exist, they can be futile in the face of STM attacks, especially within a high-stakes environment like healthcare.

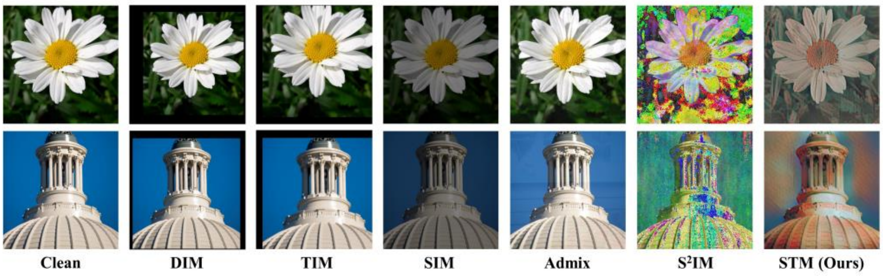


Figure 1:Sample STM Attack

3. OBJECTIVES

3.1. Main Objective

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

3.2. Sub Objective

- **Implementation of STM Attack**

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

- **Refine CNN Model**

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

- **Implementation of Defense Methods**

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

- **Evaluation of Performance Metrics**

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

- **Validate Model Defense Effectiveness**

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

4. METHODOLOGY

4.1. Overall System Diagram

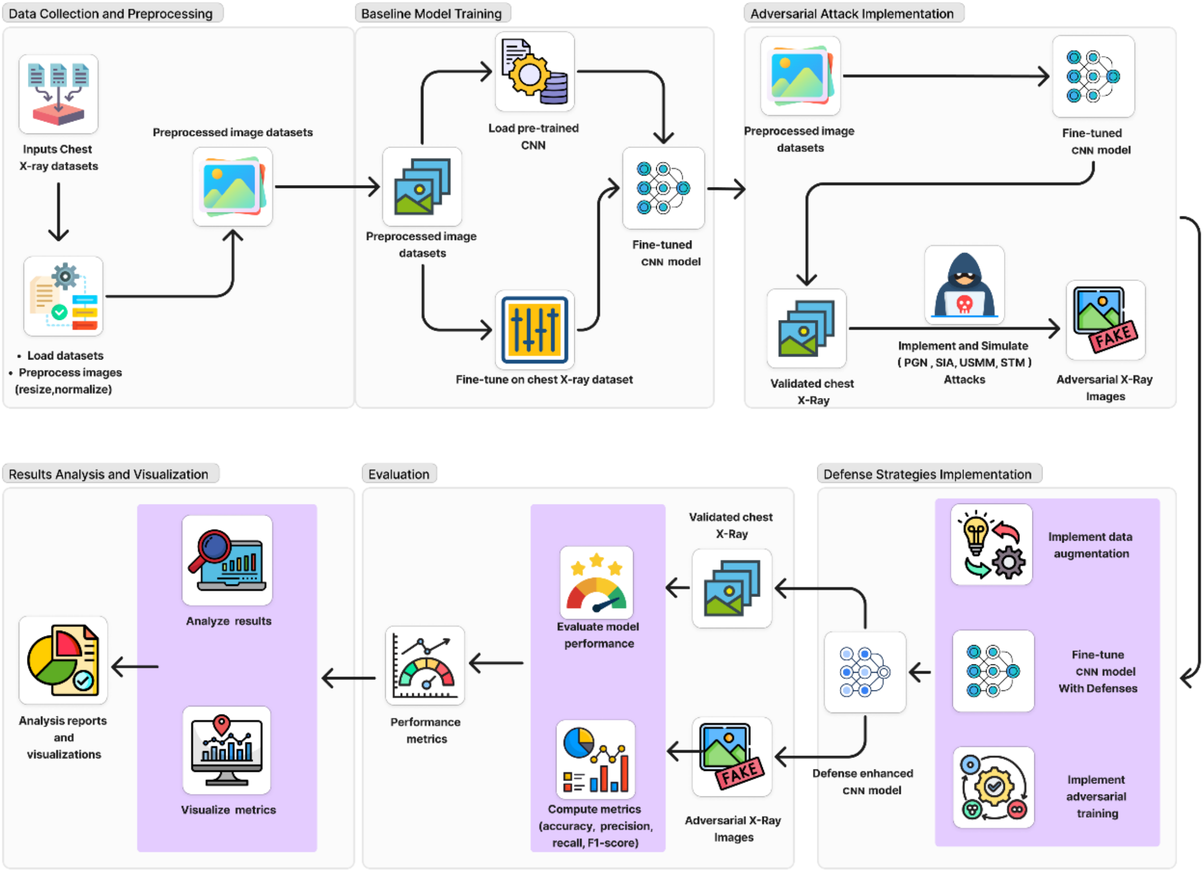

Data Collection and Preprocessing
Inputs Chest X-ray datasets
Preprocessed image datasets
• Load datasets
• Preprocess images (resize,normalize)
Baseline Model Training
Load pre-trained CNN
Preprocessed image datasets
Fine-tuned CNN model
Fine-tune on chest X-ray dataset
Adversarial Attack Implementation
Preprocessed image datasets
Fine-tuned CNN model
Validated chest X-Ray
Implement and Simulate (PGN , SIA, USMM, STM) Attacks
FAKE
Adversarial X-Ray Images
Results Analysis and Visualization
Analyze results
Analysis reports and visualizations
Visualize metrics
Evaluation
Performance metrics
Evaluate model performance
Compute metrics (accuracy, precision, recall, F1-score)
Validated chest X-Ray
FAKE
Adversarial X-Ray Images
Defense Strategies Implementation
Implement data augmentation
Fine-tune CNN model With Defenses
Implement adversarial training
Defense enhanced CNN model

Figure 2: Overall System Diagram

This diagram represents the overall workflow of your research project, starting with the collection and preprocessing of chest X-ray datasets. The process begins by gathering the necessary X-ray images, which are then resized and normalized to prepare them for model training. Next, a pre-trained CNN model is loaded and fine-tuned specifically on the chest X-ray dataset to optimize its performance for this task. Following the training, adversarial attacks, including STM and other methods like PGN, SIA, and USMM, are implemented. These attacks create adversarial X-ray images that subtly alter the texture and colors in a way that is imperceptible to the human eye but can mislead the CNN model.

To counter these attacks, various defense strategies are applied. These include data augmentation, which introduces a variety of images during training to make the model more robust, and adversarial training, where the model is trained with adversarial examples to enhance its resistance to such attacks. The CNN model is then fine-tuned again with these defenses in place. The performance of the model is evaluated using both the original and the adversarial attacked datasets, with metrics such as accuracy, precision, recall, and F1 score being calculated to assess the model's robustness. Finally, the results are analyzed and visualized, providing insights into the effectiveness of the defense strategies and helping to identify which methods offer the best protection against adversarial attacks. This comprehensive approach ensures that the CNN model remains reliable and accurate even when faced with sophisticated adversarial threats.

4.2. Subcomponent System Diagram

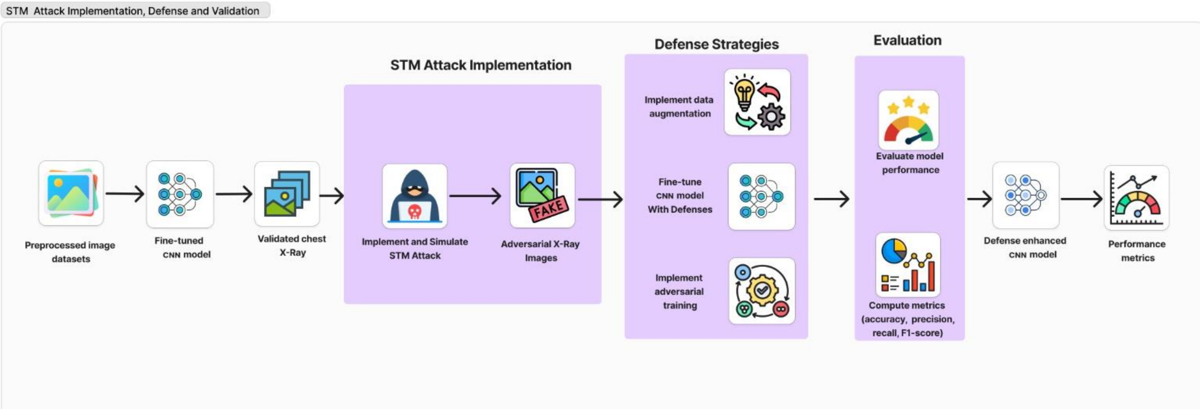

STM Attack Implementation, Defense and Validation
Preprocessed image datasets
Fine-tuned CNN model
Validated chest X-Ray
STM Attack Implementation
Implement and Simulate STM Attack
Adversarial X-Ray Images
FAKE
Defense Strategies
Implement data augmentation
Fine-tune CNN model With Defenses
Implement adversarial training
Evaluation
Evaluate model performance
Compute metrics (accuracy, precision, recall, F1-score)
Defense enhanced CNN model
Performance metrics

Figure 3: Component System Diagram

The provided system diagram outlines the key stages involved in applying Style Transfer Manipulation (STM) attacks to a CNN model,

1. Preprocessed Image Datasets

The process begins with preprocessed chest X-ray image datasets that have been resized, normalized, and prepared for training and testing the CNN model.

2. **Fine-Tuned CNN Model**

These preprocessed images are used to fine-tune an existing CNN model, optimizing it specifically for the task of chest X-ray classification. This ensures the model is well-adapted to the dataset before being subjected to adversarial attacks.

3. **Validated Chest X-Ray:**

The fine-tuned CNN model is then validated using the chest X-ray dataset to ensure its accuracy and reliability in classifying images under normal, non-adversarial conditions.

4. **STM Attack Implementation:**

The validated CNN model is subjected to STM attacks. During this stage, adversarial examples are generated by subtly altering the texture and style of the chest X-ray images, creating adversarial X-ray images that are designed to mislead the CNN model into making incorrect classifications.

5. **Defense Strategies:**

- **Data Augmentation**: Enhancing the training dataset with a wider variety of images to make the CNN model more robust against different kinds of adversarial inputs.
- **Adversarial Training**: Specifically training the CNN model with adversarial examples, enabling it to better recognize and resist STM attacks.

The CNN model is then fine-tuned again, this time incorporating these defense strategies to strengthen its resistance to adversarial attacks.

6. **Evaluation:**

- **Evaluating Model Performance:** Assessing the model's effectiveness using the adversarial attacked and clean datasets.
- **Computing Metrics:** Key performance metrics such as accuracy, precision, recall, and F1 score are calculated to determine how well the model withstands STM attacks after the defense strategies have been implemented.

4.3. Tools and Technologies

1. Deep Learning Frameworks
    - **TensorFlow:** This deep learning framework will be used to implement the CNN model and training. It supports the implementation of adversarial attacks, hence necessary research in Style Transfer Manipulation.

    - **PyTorch:** The reason for using PyTorch in this research is its dynamic computation graph and usability. It is useful for experimentation and finetuning models. It will be efficient for training and testing the CNN model against STM attacks.

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

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

4. Database Management Tools
    - **Pandas:** Pandas for data management, specifically handling chest X-ray datasets in the research. This enables a user to manipulate, clean, and analyze data efficiently. This stage in data preparation goes a long way in training models.

    - **NumPy:** NumPy will be used to enhance numerical computations through very powerful tools in array manipulation and mathematical operations. It forms an integral part in handling large datasets and computations required during model training and evaluation.

5. Visualization Tools
    - **Matplotlib:** This package will be used for visualization and plotting detailed graphs that support result analysis. It helps in visualizing the performance metrics of the model so that it would be easy to interpret how well it withstands STM attacks.

    - **Seaborn:** It is used in conjunction with Matplotlib for statistical data visualization. This library provides very high-end visualizations that help represent complex relations of data and distributions intrinsic to understanding STM attacks and their defense strategies.

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

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

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

5. PROJECT REQUIREMENTS

5.1. System Requirements

To handle the computational demands of deep learning tasks, the project requires high-performance hardware and cloud computing resources:

- **Hardware Requirements**: The system must be equipped with high-performance GPUs, such as NVIDIA Tesla or RTX, to accelerate the training and testing of CNN models. It should have a minimum of 16GB of RAM to handle large datasets and complex models, along with an SSD offering at least 500GB of storage for fast data access. A high-speed network is also essential to ensure smooth data transfer and collaboration.

- **Cloud Computing Resources**: Google Colab is recommended for cloud-based computation, offering an accessible platform for running deep learning experiments without the need for extensive local resources.

5.2. Software Requirements

The software tools and environments necessary for the project are categorized as follows:

- **Development Environment**: Python is the primary programming language, and Google Colab is suggested as the integrated development environment (IDE) due to its cloud-based capabilities, which support collaborative work and provide access to powerful computational resources.

- **Deep Learning Frameworks**: The project will utilize leading deep learning frameworks, including TensorFlow and PyTorch, which offer comprehensive tools for building and training CNN models.

- **Libraries and Tools**: Several specialized libraries are required:
    - **Adversarial Attack Libraries**: These are essential for generating adversarial examples to test the robustness of CNN models.
    - **Data Management**: Pandas and NumPy will be used for efficient data manipulation and management.
    - **Visualization**: Matplotlib and Seaborn are needed for creating detailed visual representations of data and model performance.
    - **Scikit-learn**: This library will be used for various machine learning tasks, including preprocessing, model evaluation, and implementing basic models.

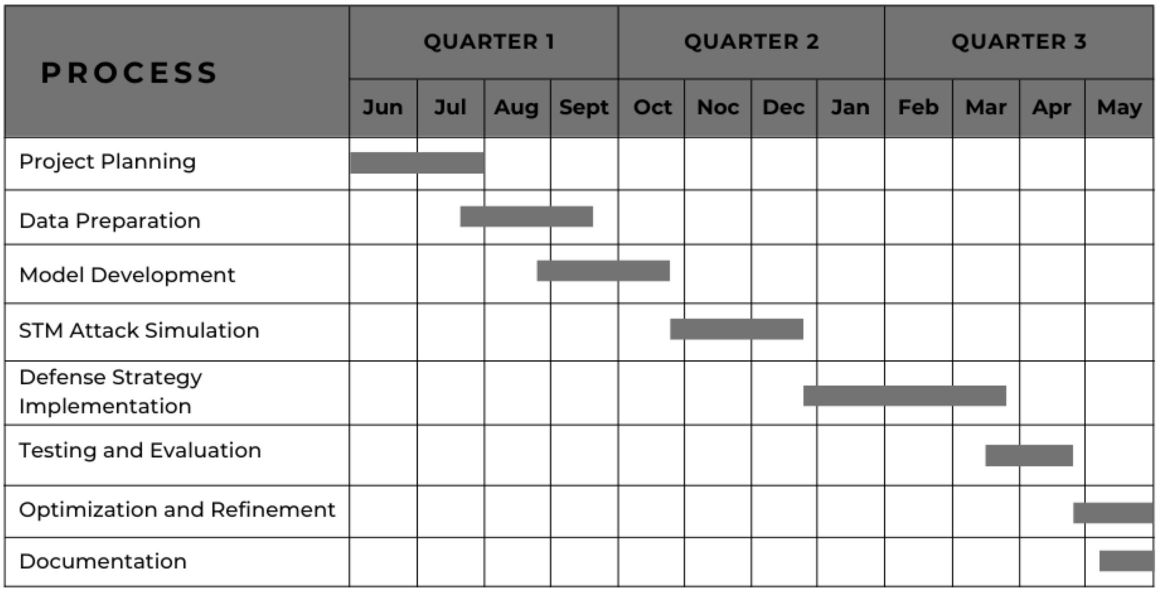
5.3. Data Requirements

- **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 both normal and abnormal cases, such as various stages of lung diseases, pneumonia, and other chest-related conditions, to ensure the model can generalize well across different scenarios.

    Data Set: https://www.kaggle.com/datasets/nih-chest-xrays/data

- **Adversarial Altered X-ray Images:** A collection of chest X-ray images that have been subjected to STM attacks is necessary. This dataset should include a variety of STM techniques applied to both normal and abnormal cases, allowing for a comprehensive evaluation of the model’s robustness against such adversarial attacks.

6. GANNT CHART



PROCESS	QUARTER 1				QUARTER 2				QUARTER 3			
	Jun	Jul	Aug	Sept	Oct	Noc	Dec	Jan	Feb	Mar	Apr	May
Project Planning	■	■										
Data Preparation		■	■	■								
Model Development			■	■	■							
STM Attack Simulation					■	■	■					
Defense Strategy Implementation							■	■	■	■		
Testing and Evaluation										■	■	
Optimization and Refinement												■
Documentation												■

Table 2: Gantt Chart

- **June to July: Project Planning**

The project begins with a comprehensive planning phase during June and July. This phase involves setting clear research goals, identifying the necessary resources, and outlining the methodologies that will be employed throughout the project. Detailed scheduling and task allocation are also part of this crucial preparatory stage.

- **July to September: Data Preparation**

As planning concludes, data preparation kicks off in July and continues through September. During this phase, essential chest X-ray datasets are collected and undergo preprocessing steps, such as cleaning, resizing, and normalization. This phase ensures that the data is in optimal condition for subsequent model training and testing.

- **August to October: Model Development**

Model development starts in August, overlapping with the final stages of data preparation, and extends through October. This involves designing and fine-tuning CNN models, focusing on optimizing them for the specific task of chest X-ray classification. This phase is critical for establishing a solid foundation for the models' performance.

- **November to January: STM Attack Simulation**

With the models developed, the project moves into the STM attack simulation phase from November to January. Here, the CNN models are subjected to STM attacks to evaluate their vulnerability. This phase is key to understanding how adversarial attacks can affect model accuracy and reliability.

- **January to March: Defense Strategy Implementation**

Defense strategies are then implemented between January and March. This phase focuses on fortifying the CNN models against the previously simulated STM attacks. Techniques such as adversarial training and data augmentation are applied to enhance the models' resilience.

- **March to April: Testing and Evaluation**

Following the implementation of defenses, the project enters the testing and evaluation phase from March to April. The performance of the defense-enhanced CNN models is rigorously assessed using both clean and adversarial datasets. Key metrics, including accuracy, precision, recall, and F1 score, are used to gauge the effectiveness of the defense strategies.

- **April to May: Optimization and Refinement**

Optimization and refinement occur from April through May. This phase involves fine-tuning the models and defense strategies based on the evaluation results, ensuring that the models achieve maximum robustness and performance.

- **May: Documentation**

The project concludes in May with the documentation phase. This final stage involves compiling the research findings, methodologies, and conclusions into a comprehensive report. The documentation ensures that the research process is thoroughly recorded and that the results are ready for dissemination or publication.

7. WORK BREAK-DOWN STRUCTURE

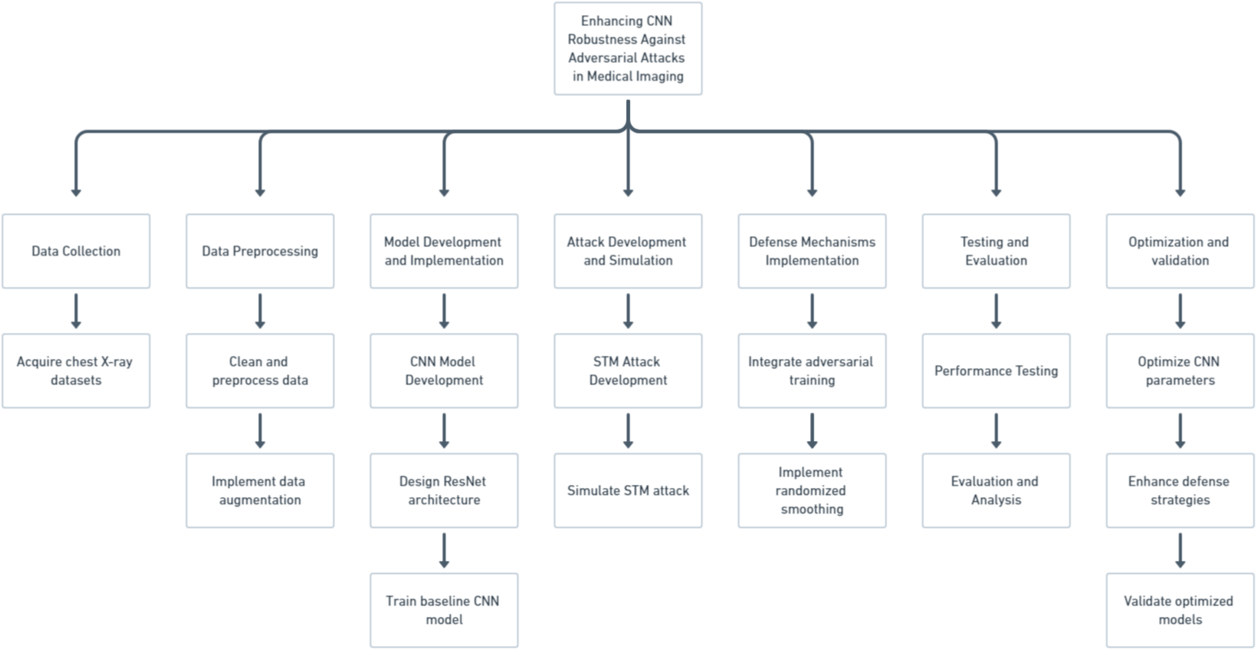

Enhancing CNN Robustness Against Adversarial Attacks in Medical Imaging
Data Collection
Acquire chest X-ray datasets
Data Preprocessing
Clean and preprocess data
Implement data augmentation
Model Development and Implementation
CNN Model Development
Design ResNet architecture
Train baseline CNN model
Attack Development and Simulation
STM Attack Development
Simulate STM attack
Defense Mechanisms Implementation
Integrate adversarial training
Implement randomized smoothing
Testing and Evaluation
Performance Testing
Evaluation and Analysis
Optimization and validation
Optimize CNN parameters
Enhance defense strategies
Validate optimized models

Figure 4: Work Breakdown Structure

Phase 1: Data Collection

- **Acquire Chest X-ray Datasets:** This project initiates with the collection of chest X-ray image datasets. The datasets would make a significant contribution to training, testing, and validation of the CNN models.

Phase 2: Data Preprocessing

- **Cleaning and Preprocessing Data:** After data collection, images are cleaned and preprocessed. This will involve resizing and normalization and making the data ready for training the model. This step is always important to ensure uniformity and quality in the dataset.
- **Implement Data Augmentation:** The dataset will be subjected to several techniques to vary the dataset in a way that will make it more robust. From here, extra images are generated through rotation, scaling, and flipping.

Phase 3: Development and Implementation of the Model

- **Development of the CNN Model:** At this stage, the CNN architecture is designed to classify chest X-rays. In the pre-processed and augmented datasets, the model will be fit to create a base in performance.
- **Training a baseline CNN model:** Using prepared datasets to obtain a baseline model to measure the intervention performance against STM attacks/defense strategies.

Phase 4: Attack Development and Simulation

- **STM Attack Development:** At this point, the focus will be on developing Style Transfer Manipulation attacks for subtly changing the texture and style but not the perceptual content of CXR images.
- **Simulate STM Attack:** The STM attacks are simulated on the trained CNN model to assess the vulnerability and understand how these attacks impact the performance of the model.

Phase 5: Implementation of Defense Mechanisms

- **Integrate Adversarial Training:** First among them is adversarial training, in which the CNN model is retrained using adversarial examples to make it more robust against attacks.
- **Apply Randomized Smoothing:** Other defense techniques, such as randomized smoothing, are used to make this CNN model even more resilient to adversarial attacks by incorporating resistance to minor perturbations in images that form the input.

Phase 6: Testing and Evaluation

- **Performance Testing:** This is for testing the performance of the defense-enhanced CNN model on clean and adversarial attacked datasets.
- **Evaluation and Analysis:** The results obtained from performance testing are used to evaluate exactly how well the defense strategies protect the CNN model. Several key metrics, such as accuracy, precision, recall, and the F1 score, are calculated to provide an all-around evaluation.

Phase 7: Optimization and Validation

- **Optimize CNN Parameters:** The CNN model and the different defense strategies are further optimized considering these evaluation results to better their effectiveness and efficiency.
- **Implement Improved Defense:** Advanced and robust defense mechanisms ensure maximum protection against STM attacks.
- **Optimal model validation:** This will involve the very last step of validating these optimized CNN models with enhanced defenses in real-world medical imaging scenarios for their robustness and reliability.

8. BUDGET AND BUDGET JUSTIFICATION

8.1. Budget

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 3: Budget

8.2. Budget Justification

- **Hardware Costs**

High-Performance GPUs: LKR 50,000 – LKR 1,00,000

This is necessary for the fast and efficient training of complex CNN models used in handling large medical imaging data sets.

RAM (Minimum 16GB): LKR 6,000 – LKR 8,000

This is required to handle large data sets, ensuring smooth operation while training or testing a model.

SSD with a minimum of 500GB: (LKR 5,000 – LKR 7,000)

This will be needed to provide adequate storage and speed for storing and quickly accessing large medical imaging files and software.

- **Cloud Computing Resources**

Google Colab Pro Subscription: (Approx. LKR 1,000 per month)

This would be required to be able to provide low-cost and scalable computing power to run deep learning experiments without incurring additional hardware costs.

- **Software costs**

Development Environment and Libraries: The development environment and libraries, such as Python and TensorFlow, will be free and open source, therefore minimizing the cost of development and giving very robust features to the project.

9. COMMERCIALIZATION

1. Identify Target Markets

To commercialize medical imaging technology, identify key markets like healthcare providers, software companies, academic institutions, and government health agencies. Analyze their needs and challenges, tailoring product development, marketing, and sales strategies to meet their specific needs, ensuring efficient and effective market entry.

2. Product Development

Identifying target markets is crucial for product development. Create tailored solutions that meet their unique demands. This can be integrated with existing systems or custom software tailored to specific segments. This ensures the product meets and exceeds the audience's expectations.

3. Demonstrating Value

To gain market trust, showcase the value of your technology through pilot projects, white papers, case studies, and participation in conferences and trade shows. These methods demonstrate its effectiveness, reliability, and potential benefits, attracting potential customers and partners.

4. Regulatory Approval

Regulatory approval is crucial in the medical field, ensuring technology meets industry standards and certifications from authorities like FDA or CE. It legitimizes the technology and opens doors to market adoption, ensuring reliability and safety for clinical use.

5. Marketing and Sales Strategy

A successful marketing and sales strategy is crucial for early product adoption. Educational marketing, direct sales, and tailoring strategies to target audience needs can help educate customers about technology benefits and applications. Tailoring these strategies helps build strong relationships and drive adoption.

6. Pricing Strategy

A strategic pricing strategy is crucial for ensuring the competitiveness and profitability of your technology. Cost-plus pricing, based on production costs plus a margin, is suitable for high-cost markets, while value-based pricing considers the perceived value of the product.

10. REFERENCES

[1] Z. Yan, C. Mao, Y. Huang, L. Zhang, and Q. Zhang, "Attacking Sequential Learning Models with Style Transfer Based Adversarial Examples," *J. Phys.: Conf. Ser.*, vol. 1880, no. 1, p. 012021, 2021. doi: 10.1088/1742-6596/1880/1/012021.

[2] W. Shen, C. Wu, Y. Xu, X. Yuan, and G. Li, "Improving the Transferability of Adversarial Examples with Arbitrary Style Transfer," *arXiv preprint arXiv:2308.10601*, Aug. 2023.

[3] S. Cheng, H. Xu, H. Zhang, M. Cheng, and N. Wang, "Towards Feature Space Adversarial Attack by Style Perturbation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 9, pp. 7534-7542, May 2021.

[4] X. Yu, T. Zhou, and C. Gao, "Towards Harmonized Regional Style Transfer and Manipulation for Facial Images," *ResearchGate*, Dec. 2022. Z

[5] A. Rana, N. P. Rana, P. Rana, P. Kumar, and R. Rana, "Chest Diseases Prediction from X-ray Images using CNN Models: A Study," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 10, no. 2, pp. 1-7, 2022.

[6] S. Narayan and A. Aggarwal, "Efficacy of Transfer Learning-based ResNet models in Chest X-ray image classification for detecting COVID-19 Pneumonia," *International Journal of Imaging Systems and Technology*, vol. 32, no. 2, pp. 454-463, June 2022. doi: 10.1016/j.imavis.2022.104033.

[7] B. Kataria, V. Kumawat, and M. Iqbal, "Deep Learning-Based Image Classification of Lungs Radiography for Detecting COVID-19 using a Deep CNN and ResNet-50," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 11, no. 1, pp. 40-48, Mar. 2023.

[8] N. Sharma, "A Comprehensive Survey on Various Deep Learning Models for COVID-19 Detection Using X-ray and CT Images," *IEEE Access*, vol. 9, pp. 112031-112049, 2021. doi: 10.1109/ACCESS.2021.3066272.

[9] S. Ahmed, M. Shahid, A. Shaikh, and R. Rana, "Comparative Analysis of Different CNN Models for COVID-19 Detection from Chest X-ray Images," *CMC: Computers, Materials & Continua*, vol. 66, no. 2, pp. 1209-1222, 2021. doi: 10.32604/cmc.2021.013232.

[10] R. Kumar and R. Acharya, "AI-driven COVID-19 detection using chest X-ray images: A comprehensive review," *Hindawi Journal of Healthcare Engineering*, vol. 2022, p. 9036457, 2022. doi: 10.1155/2022/9036457.

[11] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image Style Transfer Using Convolutional Neural Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 2414-2423.

[12] L. A. Gatys, A. S. Ecker, and M. Bethge, "A Neural Algorithm of Artistic Style," *arXiv preprint arXiv:1508.06576*, 2016.

[13] Y. Li, H. Chen, and X. Wang, "Image Classification in Medical Imaging: Security Threats and Solutions," *International Journal of Medical Informatics*, vol. 162, p. 104754, 2022.