

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

Dilsara W.N

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

Department of Computer Systems and Engineering

Sri Lanka Institute of Information Technology

Sri Lanka

August 23<sup>rd</sup>, 2024

# **ASSESSING CNN ROBUSTNESS TO UNIFORM SCALE MIX MASK (USMM) ADVERSARIAL ATTACKS**

## **Project Proposal Report**


B.Sc. (Hons) Degree in Information Technology

specialized in Cybersecurity

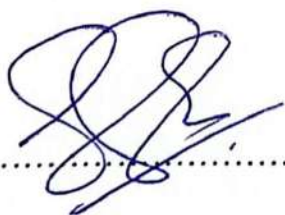
## Declaration

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
Dilsara W.N	IT21182600	

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



Signature of the Supervisor

22/8/24

Date



Signature of the Co - Supervisor

22/8/24

Date

## Abstract

The increasing reliance on Convolutional Neural Networks (CNNs) for medical imaging, particularly in chest X-ray classification, has highlighted both the potential and vulnerabilities of these models in clinical settings. This research focuses on assessing the robustness of CNN models against the Uniform Scale Mix Mask (USMM) adversarial attack, a sophisticated technique that modifies image dimensions uniformly and blends parts of other images, creating subtle but potentially harmful perturbations. These modifications, while imperceptible to the human eye, can lead to significant misclassifications by the CNN models, compromising patient safety and diagnostic accuracy. The study will implement and simulate the USMM attack on a fine-tuned CNN model, followed by an evaluation of defense strategies, including data augmentation and adversarial training, to enhance the model's resilience. The effectiveness of these defenses will be measured using key performance metrics such as accuracy, precision, recall, and F1-score under both clean and adversarial conditions. By addressing the gaps in existing research and providing a comprehensive analysis of CNN robustness against USMM attacks, this study aims to contribute to the development of more secure and reliable medical imaging systems, thereby ensuring the accuracy and trustworthiness of AI-driven diagnostics.

**Keywords:** *CNN robustness, Secure Medical Diagnostics, adversarial attacks, Uniform Scale Mix Mask (USMM), defensive strategies*

## TABLE OF CONTENT

Declaration.....	iii
Abstract.....	iv
TABLE OF CONTENT .....	v
TABLE OF FIGURES .....	vi
TABLE OF TABLES .....	vi
1 INTRODUCTION.....	1
2 BACKGROUND & LITERATURE SURVEY.....	2
2.1 Overview .....	2
2.2 Literature Survey.....	2
2.3 Research Gap .....	4
2.4 Research Problem.....	6
3 OBJECTIVES .....	8
3.1 Main Objectives .....	8
3.2 Specific Objectives.....	8
4 METHODOLOGY .....	10
4.1 Overall Approach .....	10
4.2 Subcomponent System Diagram .....	11
4.3 Tools and Technologies .....	12
5 PROJECT REQUIREMENTS .....	15
5.1 System Requirements.....	15
5.2 Personal Requirements.....	15
5.3 Software Requirements .....	15
6 GANNT CHART.....	17
7 WORK BREAK-DOWN STRUCTURE.....	20
8 BUDGET AND BUDGET JUSTIFICATION .....	21
8.1 Hardware Costs .....	21
8.2 Cloud Computing Resources.....	22
8.3 Software Costs .....	22
9 REFERENCES.....	23

**TABLE OF FIGURES**

Figure 1 : Sample USMM Attack.....7

Figure 2: Overall System Diagram.....10

Figure 3 : Component System Diagram .....12

Figure 4 : Work Breakdown Structure.....20

**TABLE OF TABLES**

Table 1 : Research Gap.....4

Table 2 : Gannt Chart .....17

# 1 INTRODUCTION

Medical imaging systems have become integral to modern healthcare, playing a crucial role in diagnosing various diseases. These systems rely heavily on advanced machine learning models, particularly Convolutional Neural Networks (CNNs), which excel in accurately identifying conditions from medical images such as X-rays, MRIs, and CT scans. For instance, CNN models like ResNet are frequently employed to interpret chest X-rays, diagnosing conditions like pneumonia, lung cancer, and tuberculosis with high precision.

However, as these models become more deeply embedded in healthcare, they also become increasingly susceptible to adversarial attacks. Such attacks can manipulate these models, altering their predictions in ways that can have potentially disastrous consequences. In an adversarial attack, even small, often imperceptible modifications to input images can lead to significant misclassifications. These errors can manifest as either false positives—where a healthy patient is wrongly diagnosed with a disease—or false negatives—where a diseased patient is incorrectly labeled as healthy. Both outcomes pose serious risks, potentially leading to incorrect treatment decisions and delayed access to appropriate care.

A particularly sophisticated and insidious form of adversarial attack is the Uniform Scale Mix Mask (USMM) attack. The USMM attack subtly alters the dimensions of an image and blends segments from different images into it, thereby creating adversarial examples that can deceive CNN models without raising suspicion. The changes introduced by a USMM attack are typically so subtle that they go unnoticed by the human eye, yet they can cause CNN models to fail in making accurate predictions. For example, when applied to medical imaging, a USMM attack might distort a chest X-ray just enough to mislead the model into missing a critical diagnosis or falsely identifying a condition that isn't actually present.

Given the vital role that CNN models play in medical diagnostics, it is imperative to ensure their robustness against such adversarial threats. This research seeks to investigate the robustness of CNN models, particularly in the classification of chest X-ray images, against USMM attacks.

## **2 BACKGROUND & LITERATURE SURVEY**

### **2.1 Overview**

This research project focuses on assessing the robustness of Convolutional Neural Networks (CNNs) in medical imaging, specifically in the context of the Uniform Scale Mix Mask (USMM) adversarial attack. The project involves implementing the USMM attack on CNN models used in chest X-ray classification, analyzing the impact of the attack on model performance, and developing strategies to mitigate its effects.

The study begins with the collection and preprocessing of medical imaging data, including resizing, normalization, and transformation to prepare the data for model training. The CNN models are then trained and fine-tuned on these datasets. After training, the USMM attack is applied to test the models' robustness. To counteract the adversarial effects, various defense mechanisms, such as data augmentation and adversarial training, are implemented. The models are then evaluated using key performance metrics like accuracy, precision, recall, and F1-score to determine their resilience against adversarial perturbations.

In summary, this research aims to improve the reliability and security of CNNs in medical imaging by addressing the vulnerabilities exposed by the USMM attack and developing effective defense strategies.

### **2.2 Literature Survey**

A comprehensive literature review was conducted to explore existing research on adversarial attacks, particularly the Uniform Scale Mix Mask (USMM) technique, within the context of machine learning models and medical imaging.

In 2023, Tao Wang et al. presented a significant advancement in adversarial attack methods with the introduction of the USMM technique. Their work, titled "Boost Adversarial Transferability by Uniform Scale and Mix Mask Method," demonstrated how the USMM attack enhances the transferability of adversarial examples by uniformly scaling and blending image segments. Although this technique was initially applied to general-purpose image datasets, its relevance to medical



imaging, where the stakes of misclassification are exceptionally high, suggests a critical area for further investigation [1].

Rana et al. (2022) conducted a study on the use of CNN models for predicting chest diseases from X-ray images. While their research highlights the effectiveness of CNNs in medical diagnostics, it does not delve into the vulnerabilities these models face from adversarial attacks like USMM. This omission underscores the need for research that addresses how such attacks could undermine the reliability of CNN-based diagnostic tools [2].

Similarly, the study by Sharma et al. (2023) focused on lung disease classification using hybrid CNN models, particularly the Inception-ResNet-v2 architecture. Despite achieving high accuracy, the study did not consider the impact of adversarial attacks, such as USMM, which could significantly affect model performance in clinical settings. This gap points to the necessity of incorporating adversarial resilience into the development of CNN models for healthcare applications [3].

Gege Qi et al. (2021) explored stabilized medical image attacks, highlighting the susceptibility of CNNs in medical imaging to various adversarial techniques. Their research illustrates the potential risks posed by adversarial attacks to diagnostic accuracy, but it does not specifically address the USMM attack, leaving a critical gap in the literature that this research aims to fill [4].

In addition, Kraidia et al. (2024) examined defense strategies against adversarial attacks, proposing robust and efficient neural network models as a solution. Although their work offers valuable insights into general defense mechanisms, it does not specifically address defenses against the USMM attack. This suggests a pressing need for targeted defense strategies that can mitigate the specific threats posed by the USMM technique in medical imaging contexts [5].

In summary, while existing literature provides a solid foundation for understanding adversarial attacks on CNNs and discusses various defense mechanisms, there remains a significant gap in research specifically focused on the USMM attack within the domain of medical imaging. This research aims to bridge this gap by applying the USMM attack 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

Despite the growing use of Convolutional Neural Networks (CNNs) in medical imaging, particularly for tasks such as chest X-ray classification, there are critical gaps in the research regarding their vulnerability to adversarial attacks. One of the most concerning yet underexplored adversarial methods is the Uniform Scale Mix Mask (USMM) attack. This attack involves subtle modifications to image dimensions and blending parts of other images, which can lead to significant misclassifications by CNN models.

	Research 1	Research 2	Research 3	Our Research
Robustness of CNNs against USMM attacks.	✗	✓	✗	✓
Impact of USMM attacks on medical image models.	✗	✗	✗	✓
USMM attacks on medical image datasets.	✗	✗	✓	✓
Basic data augmentation techniques used adversarial defences	✗	✗	✗	✓
Performance metrics under USMM attacks.	✗	✗	✗	✓

Table 1 : Research Gap

### I. Robustness of CNNs Against USMM Attacks:

**Gap:** While adversarial attacks like FGSM (Fast Gradient Sign Method) and PGD (Projected Gradient Descent) have been extensively studied, the robustness of CNN models against USMM attacks has received little attention. Given the subtlety and potential impact of USMM attacks, this gap leaves a significant uncertainty regarding the security of CNN models, particularly in high-stakes applications like medical diagnostics.

**Our Focus:** Our research aims to address this gap by evaluating the robustness of CNN models, specifically those used in chest X-ray classification, against USMM attacks. This will provide crucial insights into the vulnerabilities of these models and inform strategies to enhance their resilience.

## II. Impact of USMM Attacks on Medical Image Models:

**Gap:** There is a distinct lack of research on how USMM attacks specifically affect the performance of medical image models. Existing studies have not adequately investigated how these attacks influence critical performance metrics, such as accuracy, precision, recall, and F1-score, within the context of medical diagnostics.

**Our Focus:** This study will systematically analyze the impact of USMM attacks on CNN models used in medical imaging. By measuring changes in performance metrics, we aim to quantify the extent of performance degradation and identify the most vulnerable aspects of these models.

## III. USMM Attacks on Medical Image Datasets

**Gap:** Research on how USMM attacks impact the integrity of medical image datasets is almost non-existent. There is a critical gap in understanding how these attacks could distort data distributions or compromise the datasets used for training CNN models, which could lead to biased or inaccurate model outputs.

**Our Focus:** Our research will investigate the effects of USMM attacks on the datasets used to train and validate CNN models. We will explore how these attacks might alter the dataset and affect the overall learning process, potentially leading to compromised model performance.

## IV. Basic Data Augmentation Techniques Used in Adversarial Defenses

**Gap:** Although data augmentation is a widely used defense strategy, its effectiveness against USMM attacks has not been sufficiently tested. Most research on data augmentation has focused on defending against more common adversarial attacks, leaving a gap in understanding its role in mitigating USMM-specific threats.

**Our Focus:** This study will evaluate the efficacy of basic data augmentation techniques as a defense mechanism against USMM attacks. By testing these techniques in the context of medical imaging, we aim to determine whether they can effectively protect CNN models from USMM-induced perturbations.

## V. Performance Metrics Under USMM Attacks

**Gap:** There is a significant lack of detailed studies examining how performance metrics are affected under USMM attack conditions. Without this understanding, the full impact of USMM attacks on CNN model reliability in clinical settings remains unclear.

**Our Focus:** Our research will investigate how USMM attacks influence key performance metrics such as accuracy, precision, recall, and F1-score. By providing a detailed analysis of these effects, we aim to offer a clearer picture of the risks posed by USMM attacks and suggest strategies to maintain model performance.

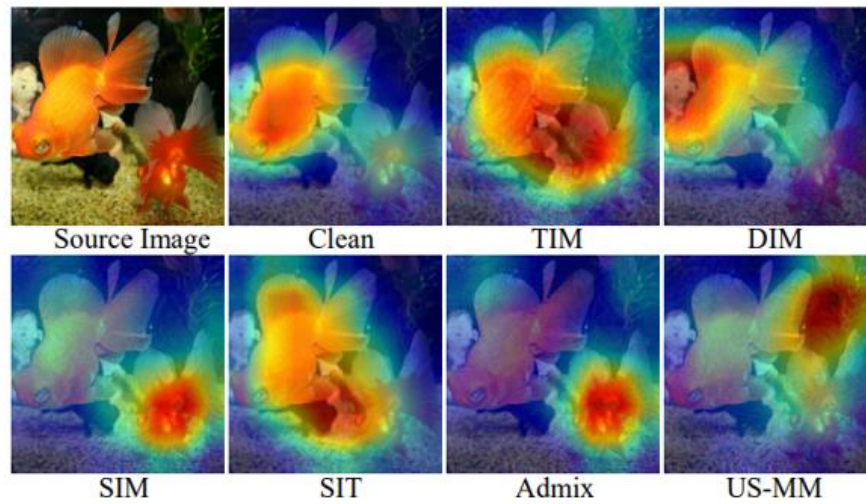
## 2.4 Research Problem

The adoption of Convolutional Neural Networks (CNNs) in medical imaging, particularly for tasks such as chest X-ray classification, has significantly improved diagnostic accuracy and efficiency. However, these models are increasingly vulnerable to adversarial attacks—malicious alterations to input data that can lead to incorrect model outputs. One such sophisticated and underexplored adversarial technique is the Uniform Scale Mix Mask (USMM) attack, which involves subtly altering image dimensions and blending segments from other images. This type of attack can cause CNN models to make significant misclassifications, thereby jeopardizing the reliability of AI-driven diagnostic tools.

Despite the critical implications of such vulnerabilities, current research lacks a comprehensive understanding of how CNN models, particularly those employed in medical imaging, respond to USMM attacks. The few existing studies on USMM attacks have not been applied within the context of medical imaging, leaving a significant gap in the literature. Furthermore, the effectiveness of existing defense strategies, such as data augmentation and adversarial training, against USMM attacks remains largely unexplored.

The specific research problem this study seeks to address is: **How vulnerable are CNN models used in chest X-ray classification to USMM adversarial attacks, and what is the effectiveness of existing defense strategies in mitigating these attacks?**

This research will focus on evaluating the impact of USMM attacks on the performance of CNN models in medical imaging and testing the resilience of these models when equipped with various defense mechanisms. The findings will provide critical insights into the vulnerabilities of CNNs in medical applications and contribute to the development of more robust and secure AI systems for healthcare.



*Figure 1 : Sample USMM Attack*

### **3 OBJECTIVES**

#### **3.1 Main Objectives**

The primary objective of this research is to thoroughly investigate the susceptibility of CNN models, particularly those utilized in chest X-ray classification, to Uniform Scale Mix Mask (USMM) attacks. These attacks involve subtle alterations to the dimensions and blending of image segments, which can mislead CNN models into making incorrect classifications, potentially compromising diagnostic accuracy. This research aims not only to 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 USMM attacks, ensuring that these models maintain their reliability and accuracy in clinical settings, even when exposed to adversarial threats.

#### **3.2 Specific Objectives**

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

##### **1. Implement the USMM Attack**

Develop and implement the Uniform Scale Mix Mask (USMM) adversarial attack on chest X-ray datasets. This includes designing the USMM attack algorithm to create perturbations that alter image dimensions and blend image segments, testing how these affect the classification performance of CNN models.

##### **2. Fine-Tune the CNN Model**

Fine-tune CNN models to optimize their performance under USMM attack conditions. This involves adjusting model parameters and training CNN on a combination of clean and USMM-perturbed data to improve its resilience against adversarial attacks.

##### **3. Test Defense Strategies**

Rigorous testing of existing defense strategies, such as data augmentation and adversarial training, to assess their effectiveness in protecting CNN models from USMM attacks. The effectiveness will be evaluated based on how well these strategies maintain model performance metrics under attack conditions.

##### **4. Evaluate Performance Metrics**

Evaluate the impact of USMM attacks on key performance metrics of CNN models, including accuracy, precision, recall, and F1-score. This will involve a detailed analysis

of how these metrics are affected by USMM perturbations, providing insights into the robustness of the models.

## **5. Validate Defense Effectiveness**

Validate the effectiveness of defense strategies by comparing the performance of CNN models with and without defenses under USMM attack conditions. The validation will help determine which strategies offer the most significant protection and maintain model reliability.

## 4 METHODOLOGY

### 4.1 Overall Approach

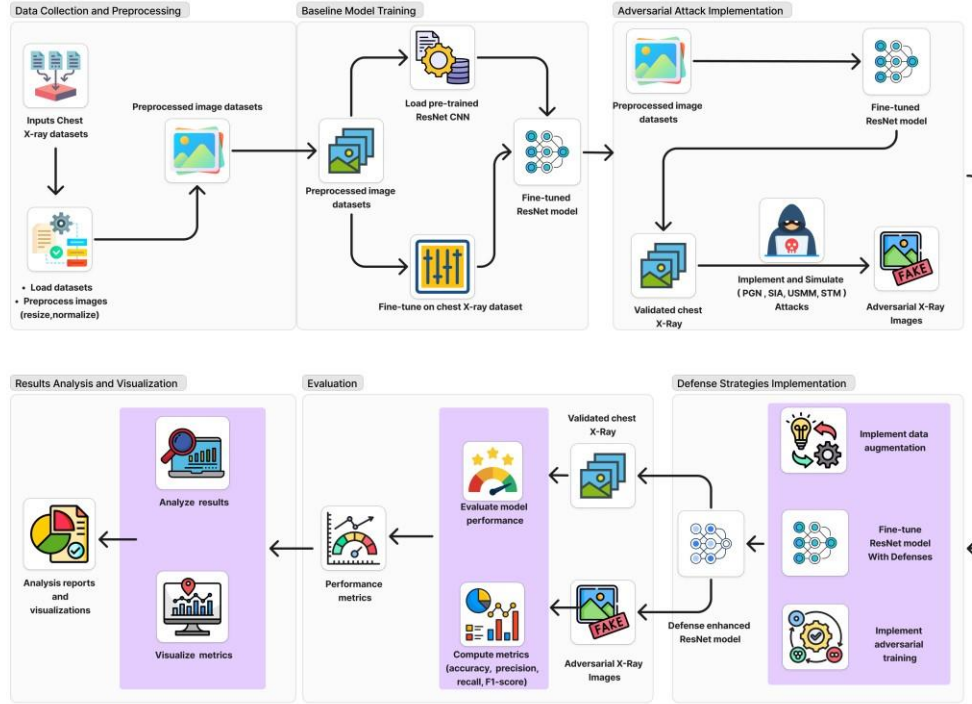


Figure 2: Overall System Diagram

This diagram illustrates the comprehensive workflow of this research project, focusing specifically on the Uniform Scale Mix Mask (USMM) adversarial attack and its impact on Convolutional Neural Networks (CNNs) used in chest X-ray classification. The methodology begins with the collection and preprocessing of chest X-ray datasets, which are essential for training and evaluating the CNN models.

The process starts by gathering chest X-ray images from publicly available medical datasets. These images undergo preprocessing steps, including resizing and normalization, to ensure uniformity across the dataset. This preprocessing is crucial to prepare the data for effective model training and evaluation.

The next step involves loading a pre-trained CNN model, which serves as the baseline for this study. The CNN model is fine-tuned on the chest X-ray dataset to optimize its performance for the specific



task of medical image classification. Fine-tuning is essential for adapting the model to the nuances of chest X-ray images, improving its accuracy and robustness.

Following the fine-tuning process, the research focuses on implementing the Uniform Scale Mix Mask (USMM) adversarial attack. The USMM attack involves subtle modifications to the image dimensions and the blending of segments from different images, creating adversarial examples that can confuse the CNN model. These adversarial images are designed to appear normal to the human eye but contain perturbations that can lead to misclassification by the model. The implementation of the USMM attack is a critical component of this research, as it tests the CNN model's vulnerability to this specific type of adversarial threat.

To mitigate the effects of the USMM attack, various defense strategies are applied. The primary defenses include data augmentation, which introduces variability into the training data by applying different transformations to the images, and adversarial training, where the model is trained on a combination of clean and adversarially perturbed images. These strategies aim to enhance the model's resistance to adversarial attacks, making it more robust and reliable in real-world applications.

After integrating the defense mechanisms, the CNN model is fine-tuned again with these defenses in place. This step is crucial to ensure that the model can effectively counter the USMM attack while maintaining high performance in classification tasks.

The performance of the CNN model is then evaluated using both the original clean datasets and the adversarially attacked datasets. Key performance metrics such as accuracy, precision, recall, and F1-score are calculated to assess the model's robustness against the USMM attack. This evaluation provides a detailed analysis of how well the defense strategies work in protecting the model from the specific adversarial threat posed by the USMM attack.

Finally, the results are analyzed and visualized to provide insights into the effectiveness of the defense strategies. The analysis helps identify the most effective methods for protecting CNN models against USMM attacks, ensuring that the models remain reliable and accurate even when faced with sophisticated adversarial challenges.

## **4.2 Subcomponent System Diagram**

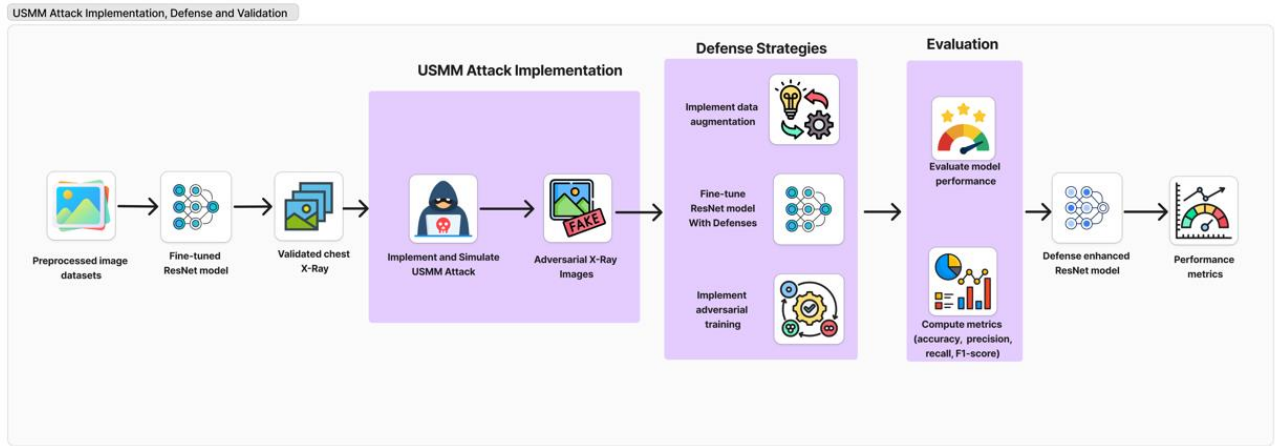


Figure 3 : Component System Diagram

## 4.3 Tools and Technologies

### 1. Deep Learning Frameworks

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

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

### 2. Programming Language

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

### 3. Model Architecture

**CNN Model:** The 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 is the package used for visualization and plotting detailed graphs which support result analysis. It helps in visualizing the performance metrics of the model so that it would be easy to interpret how well it withstands under STM attacks.

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

#### 6. Evaluation Metrics

**Scikit-learn:** This is a 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 Personal Requirements

This section details the skill set required for personnel involved in the project:

- **Skill Set:** The team members should have strong expertise in **Machine Learning and Deep Learning**, particularly in training and evaluating CNN models. Proficiency in **Python programming** and the ability to work with relevant libraries are essential for implementing algorithms and managing data. Additionally, skills in **Data Analysis** are necessary to interpret model outputs, while **Visualization** skills will help in presenting results clearly and understandably.

### 5.3 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

## 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			■	■	■							
USMM Attack Simulation					■	■	■					
Defense Strategy Implementation							■	■	■	■		
Testing and Evaluation										■	■	
Optimization and Refinement												■
Documentation												■

Table 2 : Gannt Chart

- **Project Planning (June - July):**

During this initial phase, the focus is on defining the research objectives related to the Uniform Scale Mix Mask (USMM) adversarial attack. A detailed research plan is developed, which includes identifying necessary resources, setting timelines, and outlining the methodologies to be used throughout the project. This phase ensures a clear roadmap is established to guide the subsequent research activities.

- **Data Preparation (July - September):**

This phase involves the collection and preprocessing of the chest X-ray datasets that will be used for training the CNN models. Preprocessing tasks include resizing, normalizing, and organizing the images to ensure consistency and quality across the dataset. The prepared dataset will be crucial for the accurate training and testing of the CNN models.

- **Model Development (August - October):**

The focus during this phase is on building and fine-tuning the CNN models. This includes setting up the model architecture and training the model on the prepared chest X-ray datasets. The objective is to optimize the model's performance for classification tasks, ensuring it is well-prepared to handle the complexities introduced by adversarial attacks.

- **USMM Attack Simulation (November - January):**

In this phase, the USMM attack is implemented and simulated on the trained CNN models. The USMM attack involves subtle modifications to the image dimensions and blending segments from different images to create adversarial examples. The purpose of this phase is to assess how these adversarial examples affect the CNN model's ability to accurately classify chest X-ray images.

- **Defense Strategy Implementation (January - March):**

This phase involves testing and applying various defense strategies to the CNN models. Defense mechanisms such as data augmentation and adversarial training are implemented to enhance the model's resilience against the USMM attacks. The goal is to strengthen the model's ability to resist adversarial perturbations and maintain high classification accuracy.

- **Testing and Evaluation (March - April):**

- During this phase, the CNN models are rigorously tested on both clean and adversarially attacked datasets. Key performance metrics, including accuracy, precision, recall, and F1-score, are evaluated to determine the effectiveness of the implemented defense strategies. This testing phase is critical for understanding how well the defenses protect the model from USMM attacks.

- **Optimization and Refinement (April - May):**

Based on the results from the testing and evaluation phase, the models and defense strategies are refined and optimized. The goal is to further improve the robustness and performance of



the CNN models against USMM attacks. This phase ensures that the final model is as effective and resilient as possible before final documentation.

- **Documentation (May):**

In the final phase, all research findings, methodologies, and results are thoroughly documented. This includes preparing the final report or paper that compiles the entire research process and outcomes. The documentation ensures that the research is ready for submission or publication, providing a comprehensive record of the work conducted.

There are notable gaps when it comes to the application of these findings in medical imaging, particularly in relation to the USMM adversarial attack. Current studies have not sufficiently addressed how USMM attacks can degrade the performance of CNNs used in medical diagnostics, nor have they thoroughly evaluated the efficacy of existing defense strategies in this specific context. Additionally, while some research has explored the impact of adversarial attacks on general image classification, there is limited understanding of how these attacks affect the nuanced and critical task of medical image interpretation. This research seeks to fill these gaps by providing a focused investigation into the impact of USMM attacks on CNNs in medical imaging and by rigorously testing the defense mechanisms that could protect against such threats.

## 7 WORK BREAK-DOWN STRUCTURE

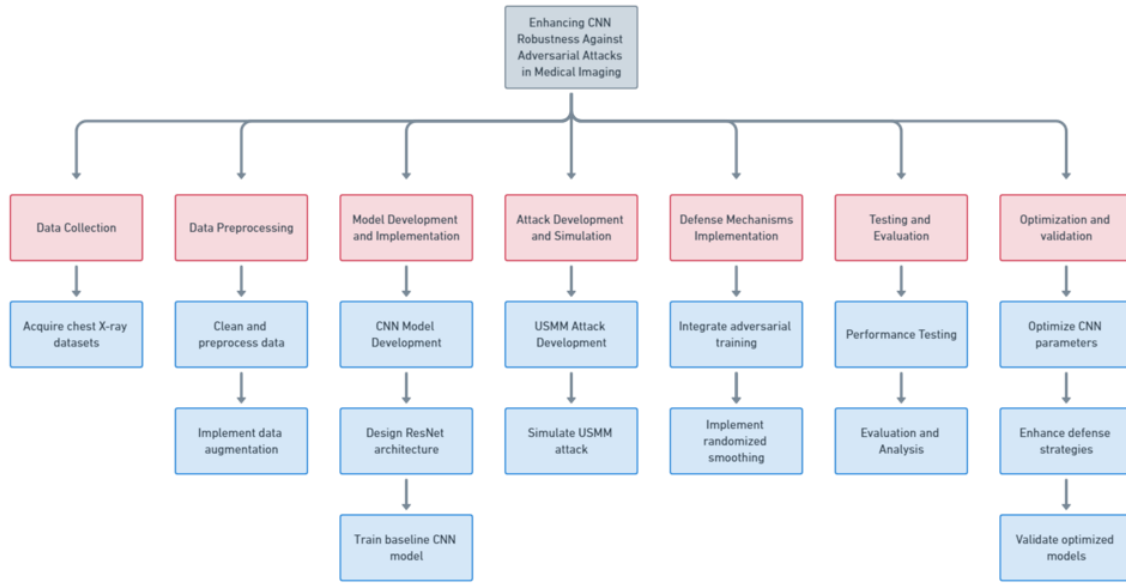


Figure 4 : Work Breakdown Structure

The project begins with a **Project Planning** phase from June to July, where the research objectives are clearly defined, necessary resources are identified, and a detailed plan and timeline are developed to guide the entire project. Following this, during the **Data Preparation** phase from July to September, chest X-ray datasets are collected, resized, normalized, and validated to ensure they are ready for model training. From August to October, the focus shifts to **Model Development**, where a pre-trained CNN model is loaded, fine-tuned on the prepared chest X-ray datasets, and validated to establish a performance baseline.

In the **USMM Attack Simulation** phase, taking place from November to January, the Uniform Scale Mix Mask (USMM) attack is developed and applied to create adversarial examples. The CNN model's vulnerability to these attacks is thoroughly evaluated. To counteract these adversarial effects, the **Defense Strategy Implementation** phase runs from January to March, involving the application of strategies such as data augmentation and adversarial training. The CNN model is fine-tuned again with these defenses in place to enhance its robustness.

During the **Testing and Evaluation** phase, spanning March to April, the model is rigorously tested on both clean and adversarial datasets. Key performance metrics are calculated, and the results are

analyzed to assess the effectiveness of the defense strategies. This is followed by an **Optimization and Refinement** phase from April to May, where the CNN model and defense strategies are optimized based on the evaluation results, ensuring the final model is as resilient and accurate as possible. The project concludes in May with the **Documentation** phase, where all research processes, methodologies, and findings are thoroughly documented and compiled into a final report, ready for submission or publication.

## 8 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.

### 8.1 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.

## **8.2 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.

## **8.3 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.

## 9 REFERENCES

- [1] Tao Wang, Zijian Ying, Qianmu Li & zhichao Lian, "Boost Adversarial Transferability by Uniform Scale and Mix Mask Method," Nanjing University of Science and Technology, 2023.
- [2] 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, p. 7, 2022.
- [3] Chandra Mani Sharma, Lakshay Goyal, Vijayaraghavan M. Chariar, and Navel Sharma, "Lung Disease Classification in CXR Images Using HybridInception-ResNet-v2 Model and Edge Computing," *Journal of Healthcare Engineering*, p. 16, 2023.
- [4] Gege Qi, Lijun Gong, Yibing Song, Kai Ma, Yefeng Zheng, "Stabilized Medical Image Attacks," arxiv, 2021.
- [5] Insaf Kraidia, Afifa Ghenai & Samir Brahim Belhaouari, "Defense against adversarial attacks: robust and efficient compressed optimized neural networks," *Scientific Reports*, p. 25, 2024.
- [6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Fully automatic CNN design with inception and ResNet blocks," *Neural Computing and Applications*, vol. 35, p. 1569, 2022.
- [7] Shuai Zhou, Chi Liu, Dayong Ye, Wanlei Zhou, Tianqing Zhu, Philip S. Yu, "Adversarial Attacks and Defenses in Deep Learning: From a Perspective of Cybersecurity," *ACM Journals*, vol. 55, no. 8, p. 39, 2022.