IT4010 – Research Project - 2024

Topic Assessment Form

Project ID: 24-25J-079

1. Topic (12 words max)

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

2. Research group the project belongs to

Computing Infrastructure (CI)

3. Research area the project belongs to

Cyber Security (CS)

4. If a continuation of a previous project:

Project ID	
Year	

5. Brief description of the research problem including references (200 – 500 words max) – references not included in word count.

Medical imaging systems, such as those used for chest X-ray classification, have become integral to modern healthcare, aiding in the diagnosis and treatment of various conditions. However, the deployment of these systems is increasingly threatened by adversarial attacks. Adversarial attacks involve small, often imperceptible perturbations to input images that can cause significant misclassification errors in machine learning models, including those based on convolutional neural networks (CNNs). This vulnerability poses a serious risk to patient safety and the reliability of medical diagnostics (Finlayson et al., 2019; Paschali et al., 2018).

Transfer learning models, such as ResNet CNN, have shown great promise in medical image classification due to their ability to leverage pre-trained weights on large datasets, thus improving performance on medical tasks (Raghu et al., 2019). Despite their success, these models are not immune to adversarial attacks. The robustness of these models must be thoroughly evaluated to ensure their safe application in clinical settings.

Several types of adversarial attacks have been identified, each exploiting different aspects of model vulnerability. For instance, the Penalizing Gradient Norm (PGN) attack alters the loss function to increase model sensitivity to small perturbations (Ge et al., 2023). The Split Image Adversarial (SIA) attack manipulates images by splitting them into blocks and applying various transformations (Wang et al., 2023). The Uniform Scale Mix Mask (USMM) attack combines uniform scaling with a mix mask from a different category, introducing novel perturbations (Wang et al., 2023). Lastly, Style Transfer Manipulation (STM) utilizes style transfer networks to shuffle and rotate image blocks, creating adversarial examples that can deceive models (Wang et al., 2023).

To counter these threats, various defense strategies need to be explored and implemented. Data augmentation, for example, can enhance the robustness of models by exposing them to a wider variety of perturbations during training (Shorten & Khoshgoftaar, 2019). Adversarial training, where models are trained on both clean and adversarial examples, has also proven effective in improving resilience (Goodfellow et al., 2015). Evaluating these defense strategies against a range of adversarial attacks will provide a comprehensive understanding of their effectiveness and highlight areas for further improvement.

The goal of this research is to systematically investigate the robustness of ResNet CNN models in the context of chest X-ray classification against the adversarial attacks. By implementing and evaluating various defense strategies, this study aims to enhance the security and reliability of medical imaging systems, thereby safeguarding patient safety and ensuring the accuracy of medical diagnostics.

References

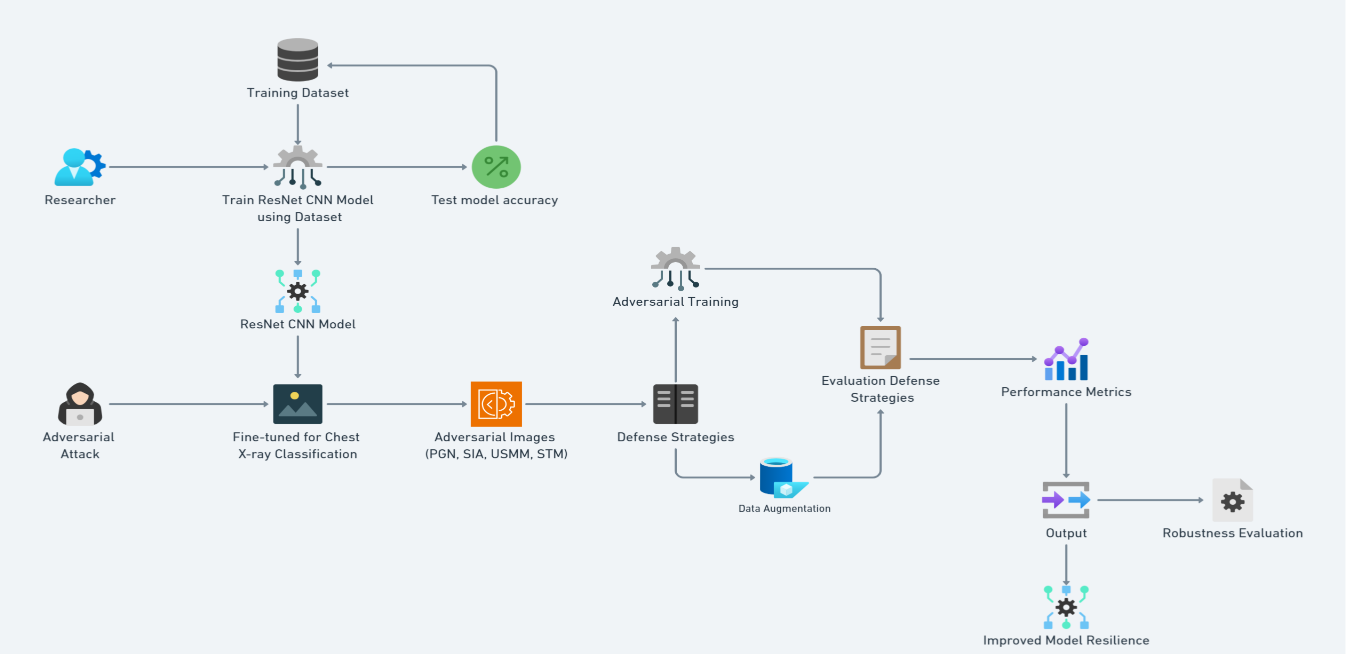
- Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). Adversarial attacks on medical machine learning. Science, 363(6433), 1287-1289.
- Paschali, M., Conjeti, S., Navarro, F., Navab, N., & Albarqouni, S. (2018). Generalizability vs. robustness: Adversarial examples for medical imaging. In MICCAI.
- Raghu, M., Zhang, C., Kleinberg, J., & Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. In NeurIPS.
- Ge, X., Wang, L., & Yang, Z. (2023). Penalizing gradient norm for robust neural networks. IEEE Transactions on Neural Networks and Learning Systems.
- Wang, J., Liu, W., & Li, X. (2023). Split Image Adversarial Attack. In CVPR.
- Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6(1), 60.
- Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In ICLR.

6. Brief description of the nature of the solution including a conceptual diagram (250 words max)

To make medical imaging systems, like those used for chest X-rays, more robust against adversarial attacks, we propose a comprehensive solution that includes both detection and defense mechanisms. First, we will create various types of adversarial attacks, such as Penalizing Gradient Norm (PGN), Split Image Adversarial (SIA), Uniform Scale Mix Mask (USMM), and Style Transfer Manipulation (STM). These attacks involve making subtle changes to the images that are hard to notice but can cause the system to make mistakes. By applying these attacks to our ResNet CNN model, which has been fine-tuned for chest X-ray classification, we can evaluate its vulnerability.

Next, we will develop defense strategies to counteract these attacks. One strategy is data augmentation, which involves creating a wide variety of training images with different changes to help the model learn to recognize real images better. Another strategy is adversarial training, where the model is trained using both normal and adversarial images to improve its resilience against these manipulations.

To assess the effectiveness of our defense strategies, we will use performance metrics such as accuracy, precision, recall, and F1-score. These metrics will help us understand how well the model performs on both clean and adversarial test sets. By systematically implementing and evaluating these components, our goal is to ensure that medical imaging systems remain reliable and secure, providing accurate diagnostics even in the face of adversarial attacks. This will ultimately safeguard patient safety and enhance the trustworthiness of medical technology.



1. Brief description of specialized domain expertise, knowledge, and data requirements (300 words max)

Specialized Domain Expertise

1. Medical Imaging: Expertise in understanding chest X-rays, and the clinical implications of accurate and inaccurate diagnoses. Familiarity with common conditions identifiable in chest X-rays, such as pneumonia, lung cancer, and tuberculosis, is essential.

2. Machine Learning and Deep Learning: Proficiency in designing, training, and fine-tuning deep learning models, especially CNNs like ResNet. Knowledge of transfer learning techniques and their application to medical imaging tasks is crucial.

3. Adversarial Machine Learning: Understanding of adversarial attacks and defenses in machine learning. Experience in implementing and evaluating adversarial attacks such as PGN, SIA, USMM, and STM is necessary.

4. Data Science and Analytics: Skills in handling large datasets, performing data augmentation, and evaluating model performance using metrics such as accuracy, precision, recall, and F1-score.

Knowledge Requirements

1. Healthcare and Radiology: Understanding the healthcare domain, particularly radiology, to comprehend the significance of robust and accurate medical imaging systems. Knowledge of how diagnostic errors can impact patient outcomes.

2. Algorithm Development: Proficiency in developing algorithms for both adversarial attacks and defense strategies. Familiarity with Python and machine learning libraries such as TensorFlow, PyTorch, and Keras.

3. Security in AI: Knowledge of security measures in AI and machine learning to develop robust defense strategies against adversarial attacks.

Data Requirements

1. Chest X-ray Datasets:

- [Labeled Chest X-ray Images] (https://www.kaggle.com/datasets/tolgadincer/labeled-chest-xray-images)
- [NIH Chest X-rays] (https://www.kaggle.com/datasets/nih-chest-xrays/data)
- [ChestX-ray14] (https://www.v7labs.com/open-datasets/chestx-ray14)
- [Chest X-ray Pneumonia] (https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia/code)
- [Chest X-ray Dataset] (https://www.kaggle.com/datasets/alifrahman/chestxraydataset)

2. Objectives and Novelty

Main Objective

Investigate the robustness of transfer learning models, specifically ResNet CNN, for medical image classification against various adversarial attacks, and develop and evaluate effective defense strategies to enhance model resilience, ensuring reliable and secure medical diagnostics.

Member Name	Sub Objective	Tasks	Novelty
N Premakanthan	Investigate the robustness of ResNet CNN for chest X-ray classification against the STM (Style Transfer and Manipulation) adversarial attack and evaluate the effectiveness of various defense strategies.	• Develop and apply STM attack using style transfer networks. • Randomly shuffle and rotate image blocks to create adversarial examples. • Fine-tune ResNet CNN and apply STM. • Develop and implement defense strategies such as data augmentation and adversarial training. • Test the resilience of ResNet CNN against STM adversarial attacks.	Utilizing style transfer techniques for generating adversarial examples and investigating the combined effectiveness of these techniques and defense strategies in medical image security.

D.S.C. Wijesuriya	Investigate the robustness of ResNet CNN for chest X-ray classification against the PGN (Penalizing Gradient Norm) adversarial attack and evaluate the effectiveness of various defense strategies.	• Implement the PGN attack on the chest X-ray dataset. • Fine-tune ResNet CNN and apply PGN. • Develop and implement defense strategies such as data augmentation and adversarial training. • Evaluate the model's performance metrics (accuracy, precision, recall, F1-score) on both clean and adversarial datasets.	Exploring the combined effectiveness of gradient norm penalization and defense strategies in enhancing model robustness specifically in the context of medical imaging.
P.G.E.J. Sandamal	Investigate the robustness of ResNet CNN for chest X-ray classification against the SIA (Split Image Adversarial Attack) adversarial attack and evaluate the effectiveness of various defense strategies.	• Implement the SIA attack by splitting the image into blocks and applying various transformations. • Fine-tune ResNet CNN and apply SIA. • Develop and implement defense strategies such as data augmentation and adversarial training. • Measure the performance of the fine-tuned ResNet CNN under SIA.	Investigating the unique approach of block-wise transformations in adversarial attacks and their specific effects on medical image classification.

W.N. Dilsara	Investigate the robustness of ResNet CNN for chest X-ray classification against the USMM (Uniform Scale Mix Mask) adversarial attack and evaluate the effectiveness of various defense strategies.	• Implement the USMM attack, integrating uniform scaling and mix masking techniques. • Fine-tune ResNet CNN and apply USMM. • Develop and implement defense strategies such as data augmentation and adversarial training. • Evaluate the ResNet CNN's performance under USMM conditions.	Introducing a combination of uniform scaling and mix masking as an adversarial strategy and exploring the combined effectiveness of this strategy and defense mechanisms for medical imaging.

3. Supervisor checklist

a) Does the chosen research topic possess a comprehensive scope suitable for a final-year project?

Yes	✓	No	

b) Does the proposed topic exhibit novelty?

Yes	✓	No	

c) Do you believe they have the capability to successfully execute the proposed project?

Yes	✓	No	

d) Do the proposed sub-objectives reflect the students' areas of specialization?

Yes	✓	No	

e) Supervisor's Evaluation and Recommendation for the Research topic:

OK.

4. Supervisor details

	Title	First Name	Last Name	Signature
Supervisor	Dr.	Harinda	Fernando	
Co-Supervisor	Mr.	Kavinga	Abeywardena	for 24/04/2024
External Supervisor				
Summary of external supervisor's (If any) experience and expertise				

This part is to be filled by the Topic Screening Panel members.

Acceptable: Mark/Select as necessary

Topic Assessment Accepted	
Topic Assessment Accepted with minor changes (should be followed up by the supervisor)*	
Topic Assessment to be Resubmitted with major changes*	
Topic Assessment Rejected. Topic must be changed	

* Detailed comments given below

Comments

The Review Panel Details

Member's Name	Signature

***Important**:

1. According to the comments given by the panel, make the necessary modifications and get the approval by the **Supervisor** or the **Same Panel**.
2. If the project topic is rejected, identify a new topic, and follow the same procedure until the topic is approved by the assessment panel.