Project Proposal

Team Members:

Xiang Siqi 1004875, Luah Shi Hui 1005512, Liu Yu 1005621

Title:

Survey: Adversarial Attacks and Defenses on Medical Imaging Classification Models

Introduction:

With the growth of AI models and applications, it becomes a critical topic to explore their robustness and security. Particularly in the healthcare sector, where AI-driven diagnostic tools and medical imaging classification models are increasingly relied upon, the integrity and reliability of these systems are paramount. Adversarial attacks, wherein slight but maliciously designed modifications to input data can deceive AI models into making incorrect predictions, pose a significant threat to the application of AI in medicine. Such vulnerabilities not only compromise patient care but also raise substantial ethical and safety concerns.

The goal of this project is to provide a comprehensive overview of adversarial attacks targeting deep neural network models on medical imaging classification tasks, the defense mechanisms developed to counter these attacks, and the challenges and opportunities for future research in this domain.

Dataset:

We plan to use the NIH Chest X-rays Dataset. It is comprised of 112,120 X-ray images with disease labels from 30,805 unique patients.

Expected inputs and outputs:

Inputs are X-ray images provided by NIH Chest X-rays Dataset. The outputs are diseases identified by the model. Moreover, we will make efforts to analyze the influence of adversarial samples on the trained model. This analysis will provide insights into the vulnerability of medical imaging AI to adversarial manipulation and the effectiveness of defense strategies.

Model Architecture:

We plan to use Convolutional Neural Networks as the baseline model. Given the survey nature of our project, we might introduce more models during the progress of the project.

Deliverables and Outcomes:

A comprehensive survey evaluating the landscape of adversarial threats and defense mechanisms in the context of deep learning models used for medical imaging classification tasks. Additionally, the survey will examine the effectiveness of some defense mechanisms, analyzing their strengths and weaknesses through various metrics such as robustness, accuracy, and computational efficiency.