Fortifying Deep Learning Models Against Adversarial Threats using Adversarial Training

Soujanya Menasagi[1], Savita Sidnal[1], Ajinkya Kulkarni[1], Sarvesh S Sanikop [1], Uday Kulkarni[1], and Vijayalakshmi M [1]

[1]School of Computer Science and Engineering, KLE Technological University, Hubballi, India
01fe22bcs225@kletech.ac.in,01fe22bcs225@kletech.ac.in,01fe22bcs225@kletech.ac.in ,01fe22bcs225@kletech.ac.in, uday_kulkarni@kletech.ac.in,mam@gmail

Abstract. Deep learning models have revolutionized critical fields like healthcare and autonomous systems but remain vulnerable to adversarial attacks, where minor perturbations in input data cause incorrect predictions. This project addresses these challenges by implementing adversarial training, a method that strengthens model robustness by incorporating adversarial examples crafted using FGSM, PGD, and CW attack into the training process. The results show that adversarially trained models achieve significant resilience, maintaining high accuracy on both clean and adversarial datasets, the ResNet50 model achieved 93.12% accuracy on clean images before adversarial training, and after adversarial training, the accuracy increased to 95.0% and the model's performance improved from 43.29% to 93.6% under adversarial attacks. This approach ensures the reliability and robustness of deep learning models in safety-critical applications while maintaining computational efficiency.

Keywords: Adversarial Attacks,Adversarial Training, CW, FGSM, Perturbations, PGD

1 Introduction

Machine learning (ML) models have achieved outstanding success in fields like healthcare, autonomous systems, and cyber security. However, their vulnerability to adversarial attacks poses a significant challenge to their safe deployment. Adversarial examples, created by introducing small, imperceptible perturbations to input data, can deceive even advanced ML models into making incorrect predictions with high confidence. These vulnerabilities are particularly concerning in critical applications where safety and reliability are crucial.Adversarial attacks can lead to severe consequences, such as diagnostic errors in healthcare or the misinterpretation of traffic signs in autonomous vehicles, putting lives at risk. Addressing these vulnerabilities has become a pressing need. While researchers have made progress in understanding adversarial attacks, the rapid evolution of attack methods and diverse ML architectures necessitate a systematic evaluation of both attack strategies and defense mechanisms.

This paper aims to address these challenges by implementing adversarial attack methods like FGSM, PGD, and CW to generate adversarial examples and evaluate their impact on ML models. It also explores defense mechanisms, such as adversarial training and robust optimization, to counter these attacks and enhance model robustness. These approaches are critical for improving the resilience of ML systems in real-world applications.

A key focus of the research is to evaluate and compare the robustness of different ML architectures under various adversarial scenarios. By analyzing their performance, this study identifies strengths and weaknesses in existing systems, providing insight for designing more secure and reliable AI models.Ultimately, this work contributes to advancing knowledge on adversarial robustness by bridging theory and practice. It aims to foster trust in AI systems by addressing adversarial challenges and enabling their safe integration into critical real-world applications. This research paves the way for more robust and secure ML systems capable of operating reliably in high-stakes environments.

2 Background Study

Deep learning models have significantly advanced various domains such as healthcare, cybersecurity, and autonomous systems due to their exceptional ability to extract complex features from large datasets. However, despite their success, these models are highly vulnerable to adversarial attacks, carefully crafted perturbations that are often imperceptible to humans but can lead to incorrect model predictions with high confidence. This vulnerability raises major concerns, particularly in safety-critical applications like medical diagnosis and autonomous navigation, where errors could have serious consequences.

Adversarial attacks exploit the sensitivity of deep neural networks to small input changes. Notable attack methods include the Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and the Carlini & Wagner (CW) attack, all of which generate adversarial examples capable of misleading even well-trained models. These methods have been widely studied and demonstrate the urgent need for effective defense mechanisms.

Among the most promising defenses is adversarial training, which involves augmenting the training dataset with adversarial examples. This approach enables the model to learn more robust decision boundaries, thereby improving its performance against attacks. By training on both clean and adversarial data, the model becomes more resilient to perturbations introduced at inference time.

3 Proposed Work

The proposed system Figure 1 defends deep learning models against adversarial attacks through a structured methodology involving three primary phases: crafting adversarial examples, performing adversarial training, and evaluating the model's robustness. Adversarial samples are generated using methods like FGSM, PGD, and CW, then used to train the model alongside clean data. This

improves the model's ability to withstand attacks. Robustness is evaluated using accuracy and performance metrics under various perturbations.

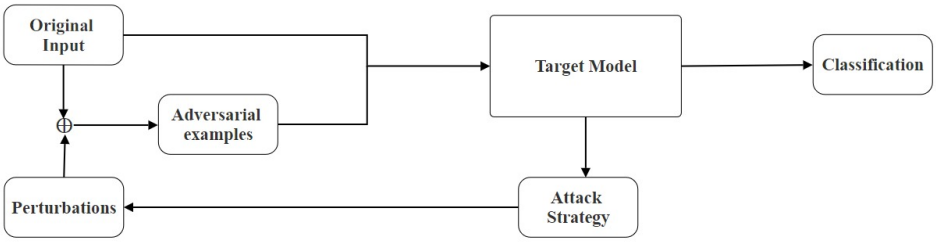


Figure 1. The Proposed System

3.1 Dataset and Preprocessing

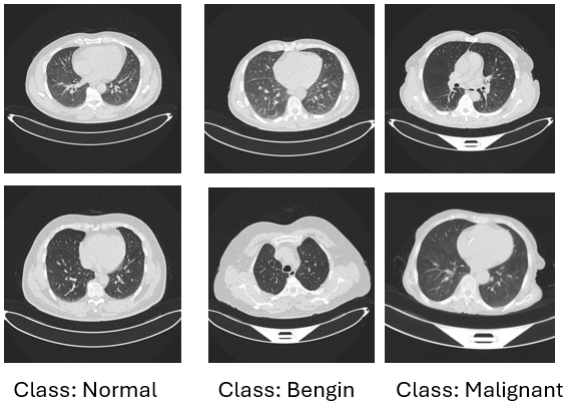


Figure 2. Sample Images from the IQ-OTH/NCCD Lung Cancer Dataset

The IQ-OTH/NCCD Lung Cancer Dataset, used in this project, contains 1,094 grayscale CT scan images categorized into three classes: benign, malignant, and normal. Benign and malignant images represent non-cancerous and cancerous conditions, respectively, while normal images indicate healthy lungs. The dataset is publicly available on Kaggle[1] and is widely used for lung cancer detection tasks. All images were preprocessed for consistency and split into training, validation, and test sets with a balanced class distribution. Figure 2 illustrates sample images from each category.

[1] `https://www.kaggle.com/datasets/hamdallak/the-iqothnccd-lung-cancer-dataset`

3.2 Adversarial Example Generation

Adversarial examples are crafted by introducing perturbations into clean input data to deceive the model into making incorrect predictions. Three methods are used for generating these examples:

Fast Gradient Sign Method (FGSM): A single-step gradient-based attack that perturbs the input data in the direction of the gradient to maximize loss. Projected Gradient Descent (PGD): An iterative attack method that applies multiple small perturbations while keeping the modified data within a predefined boundary. Carlini & Wagner (CW) Attack: An optimization-based attack that minimizes the perturbation while maximizing the probability of misclassification.

3.3 Adversarial Training

Adversarial training involves training the model on a dataset comprising both clean and adversarial examples. This improves the model's robustness by teaching it to recognize and resist adversarial perturbations. The process is as follows:

Generate adversarial examples using the FGSM, PGD, and CW methods. Combine clean and adversarial datasets. Train the model iteratively, ensuring exposure to a variety of adversarial inputs during each epoch.

3.4 Evaluation

The robustness of the trained model is evaluated by testing it against adversarial examples generated using different methods. Key metrics include accuracy on clean and adversarial datasets, robustness scores, and confusion matrices.

4 Results and Discussion

This section presents the results of the proposed system, including its performance in adversarial and clean scenarios, and discusses the outcomes of the adversarial training process. The results are interpreted using snapshots of key processes and metrics, providing insights into the robustness of the system.

4.1 Results of Normal Training

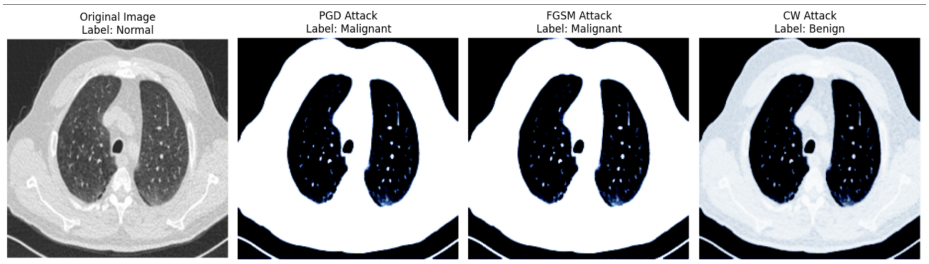


Figure 3. Model Performance During Normal Training For Normal Lung Image

Initially, the neural network model was trained on a clean dataset without any adversarial examples. The results, as shown in Figure 3, demonstrate accurate predictions for clean test data. However, the model performed poorly when tested against adversarial examples, highlighting its vulnerability to adversarial attacks.

4.2 Adversarial Attack Results

Adversarial examples were generated using the attack model, as illustrated in Figure 4. These adversarial images were visually similar to the original inputs but effectively misled the model into making incorrect predictions. The attack success rate was significantly high, emphasizing the need for adversarial training.

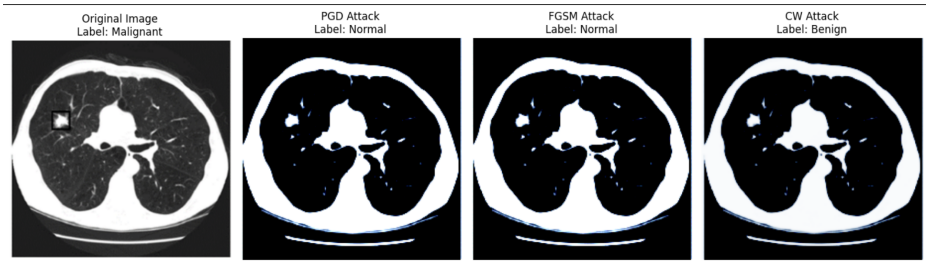


Figure 4. Model Performance During Normal Training For Malignant lung image

4.3 Results of Adversarial Training

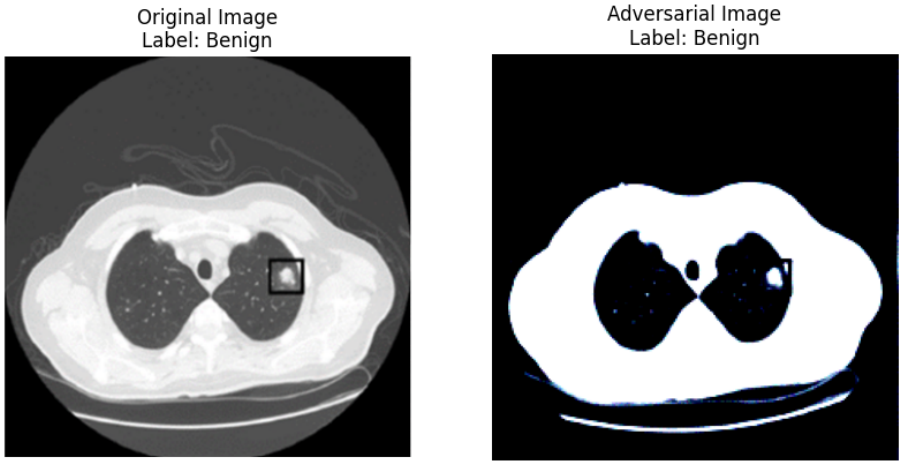
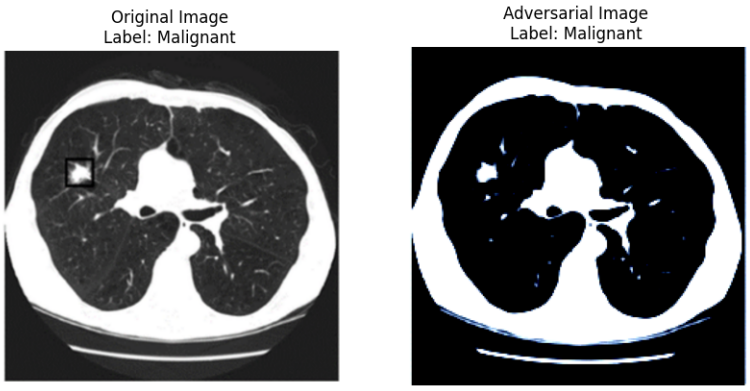


Figure 5. Prediction of Attacked Bengin image

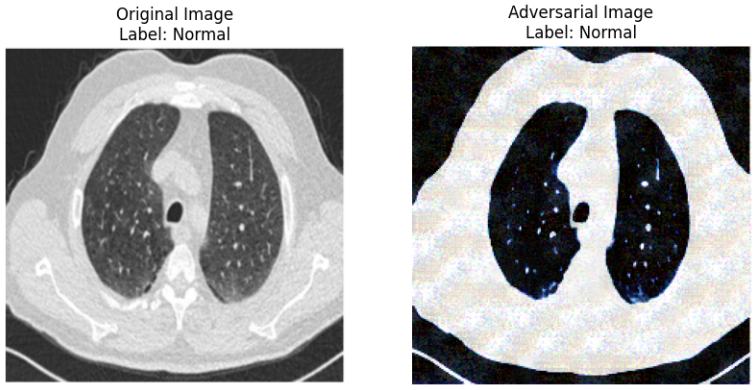
The adversarially trained model was evaluated on both clean and adversarial test datasets. As seen in Figure 5, the model achieved high accuracy on clean

data and demonstrated significant improvement in robustness against adversarial attacks. This indicates the effectiveness of the adversarial training process.

4.4 Discussion



(a) Prediction of Attacked Malignant Image



(b) Prediction of Attacked Normal Image

Figure 6. Comparison of Attacked Malignant and Normal Images

The results highlight the following key findings:

The model trained on clean data was highly accurate for clean inputs but vulnerable to adversarial examples. Adversarial examples generated by the attack model were effective in misleading the neural network, underscoring the importance of robust training techniques. Adversarial training significantly enhanced the model's robustness, achieving a balance between accuracy on clean data and resistance to adversarial perturbations. The final model demonstrated consistent performance across various unseen clean and adversarial test scenarios.

4.5 Summary of Results

The adversarial training process successfully mitigated the impact of adversarial examples, ensuring robust model performance. These results validate the proposed approach, demonstrating its potential in improving model reliability in adversarial settings.

Table 1: Accuracy on Test dataset after adversarial training

Model Training	Clean	PGD	FGSM	CW	Combined
	Image	Attacked Image	Attacked Image	Attacked Image	Adv Images
Normal Training	93.12	43.08	45.9	40.9	43.29
Adversarial Training	95.0	93.2	95.0	92.8	93.6

5 Conclusion and Future Scope

The project developed a robust framework to enhance neural network resilience against adversarial attacks through data preprocessing, adversarial example generation, and adversarial training. This approach significantly improved performance on both clean and adversarial data while maintaining high accuracy. Future work can focus on integrating real-time detection, exploring transferability to other models or datasets, and incorporating explainable AI techniques to better understand and strengthen the model's defense mechanisms.