Title: Enhancing Resilience in Medical AI: Strategies for Counteracting Adversarial Attacks in Breast Cancer Detection Models

SAIDA BINTA ALAM TAZRIF YAMSHIT RAIM SAFUAN ALAM SAIFEE ARAFATH HOSSEN SANTO

20-43880-2 21-45012-2 20-41861-1 20-43859-2

Department of Computer Sciences, American International University-Bangladesh

Abstract

This paper investigates the vulnerability of a breast cancer detection model, Model A, developed using Densenet121 and transfer learning, to adversarial attacks. Upon exposing Model, A to adversarial manipulated images using the Fast Gradient Sign Method (FGSM), its accuracy plummeted from 96% to 25%. To counter this, we developed Model B, integrating both original and adversarial altered images in the training process. This defense strategy improved the model's robustness, boosting its accuracy to 75%. The findings underscore the significance of incorporating adversarial training in medical image analysis models to ensure reliability and robustness.

Keywords: ***Adversarial Attacks, Breast Cancer Detection, Deep Learning, Transfer Learning, Densenet121, FGSM), Medical Imaging AI, Adversarial Training, AI Reliability in Healthcare, CNN, Diagnostic Accuracy, AI Vulnerabilities, Healthcare Cybersecurity***

Introduction

1. Background

Worldwide, breast cancer remains a significant medical concern, consistently listed among the top causes of cancer-related deaths in females. The paramount importance of early detection and effective treatment of breast cancer for improving patient survival is globally recognized. In this domain, the advancement of Artificial Intelligence (AI), particularly in the area of deep learning, has substantially transformed the techniques used for identifying breast cancer.AI platforms, rigorously educated on comprehensive arrays of medical images, have exhibited remarkable skill in aiding radiologists to precisely pinpoint malignant growths. Nevertheless, the integration of AI into such vital healthcare sectors demands a comprehensive examination of its inherent limitations and the development of robust strategies to address these potential drawbacks [1].

2. Problem Statement

Despite the considerable progress in integrating AI into medical imaging, the susceptibility of these models to adversarial attacks continues to pose a substantial challenge. Characterized by minor modifications in medical images, these attacks can lead deep learning system astray, resulting in erroneous diagnostic outcomes. This vulnerability critically undermines the trustworthiness of AI-driven diagnostic tools within healthcare, especially in crucial domains like cancer detection. The impact of such security breaches on patient treatment and the clinical decision-making process highlights the urgent necessity for strong defensive strategies to protect against these threats [2].

3. Objective

The focus of this study is to address the critical issue of adversarial vulnerabilities in a breast cancer detection model. We explore how susceptible Model A, which is developed using Densenet121 and transfer learning techniques, is to adversarial attacks, particularly those executed using the Fast Gradient Sign Method (FGSM). Our primary aim is to augment the robustness of this model against such attacks, ensuring its diagnostic reliability while maintaining its accuracy.

4. Contribution:

Our work makes significant contributions in two key areas. Firstly, we present empirical evidence demonstrating the vulnerability of a high-performing breast cancer detection model to adversarial attacks. Secondly, we propose and evaluate an innovative defense strategy, involving the development of Model B. This new model is trained on a dataset encompassing both original and adversarial modified images, aiming to create a system that is not only accurate in its diagnostic capabilities but also resilient to adversarial manipulations. This approach offers valuable contributions to the field of medical AI, providing insights into creating more robust and reliable AI systems for healthcare applications [3].

Related Work

In the realm of medical imaging, AI in healthcare, and adversarial learning, particularly concerning breast cancer detection, several key studies have made significant contributions. Dhasarathan et al. (2021) emphasized the need for data privacy and security in healthcare AI applications, proposing a bio-inspired privacy-preserving framework, crucial for sensitive domains like healthcare [3]. Goyal et al. (2019) demonstrated the practical application of deep learning in diagnosing skin cancer, suggesting potential parallels in breast cancer detection [4]. The integration of radiomics with machine and deep learning methods, as explored by Avanzo et al. (2020) and others, marks a crucial advancement in extracting more meaningful information from medical images for more accurate and robust AI models in medical diagnostics [5]. Parekh and Jacobs (2019) discussed the intersection of deep learning and radiomics in precision medicine, underscoring the significance of these technologies in personalizing patient care [6]. Studies by Valdora et al. (2018) and Crivelli et al. (2018) provided comprehensive insights into the application of radiomics in breast cancer, enhancing the detection and classification capabilities of AI models [7], [8]. Additionally, Li et al.'s exploration of MR imaging radiomics signatures for predicting breast cancer recurrence adds depth to the prognostic modeling in breast cancer, informing the development of more sophisticated AI detection and prediction models [9]. Collectively, these studies paint a broad picture of the current state of AI in medical imaging, highlighting the challenges, opportunities, and the necessity for security, precision, and tailored approaches in AI applications in healthcare.

Methodology

1. Dataset Overview:

For this project, we utilized the BreakHis dataset [10], a renowned and publicly available collection of breast cancer pathology images. This dataset, essential for non-commercial research on breast cancer, comprises 7902 images, including 2480 benign and 5429 malignant breast tumor images. For the efficient training, validation, and testing of our deep learning models, as well as for carrying out adversarial attack and defense experiments, we segmented the dataset into three distinct subsets:

- **Training Set:** Consisting of 1984 benign and 4343 malignant images, used for initial training.
- **Validation Set:** Comprising 248 benign and 542 malignant images, used for tuning the hyperparameters.
- **Test Set:** Including 248 benign and 544 malignant images, used for evaluating model performance.

Dataset	**Training Set**	**Validation Set**	**Test Set**	**Total**
Benign	1984	248	248	2480
Malignant	4343	542	544	5429

2. Model A (Detection Model):

Architecture: Model A is based on the Densenet121 architecture. This convolutional neural network is optimized for image classification tasks, with a particular aptitude for medical image analysis due to its capability to discern fine details, a critical factor in identifying malignancies in breast cancer images.

Training Process: We trained Model A using the annotated images from the BreakHis dataset. The training aimed to fine-tune the pre-trained Densenet121 model to enhance its accuracy in breast cancer detection. Emphasis was placed on achieving high diagnostic accuracy, with a focus on the sensitivity and specificity of the model.

3. Adversarial Attack Generation:

Technique Used: To assess the resilience of Model A, we implemented the Fast Gradient Sign Method (FGSM). This technique generates adversarial images by introducing minimal yet effective perturbations based on the model's loss function gradients relative to the input image.

Impact on Model A: Exposure of Model A to these adversarially altered images resulted in a marked decrease in diagnostic accuracy, underscoring the model's vulnerability to such attacks.

4. Model B (Defense Model):

Training with Merged Dataset: In response to Model A's vulnerabilities, we developed Model B. This model underwent training on a merged dataset, incorporating both original and FGSM-generated adversarial images. This approach aimed to acclimate the model to potential perturbations, thereby bolstering its defense against adversarial attacks. The training methodology paralleled that of Model A but included additional measures to differentiate between authentic and manipulated images.

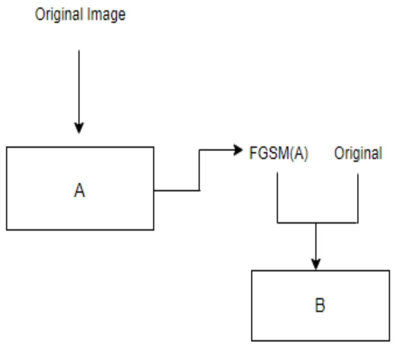

Figure 1: Model Training

5. Diagram of Proposed Model Workflow:

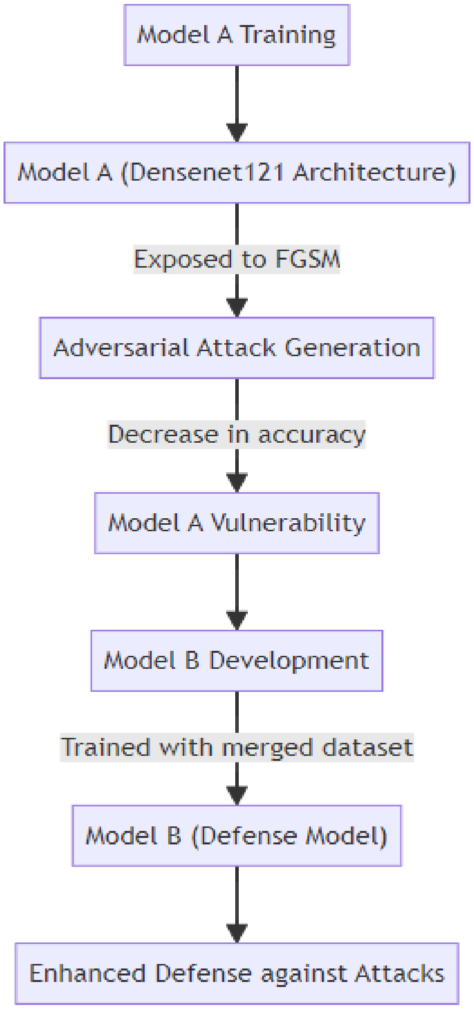
Model A Training
Model A (Densenet121 Architecture)
Exposed to FGSM
Adversarial Attack Generation
Decrease in accuracy
Model A Vulnerability
Model B Development
Trained with merged dataset
Model B (Defense Model)
Enhanced Defense against Attacks

Figure 2: Proposed Model Workflow

A comprehensive diagram that illustrates the sequential process from Model A's training, through the generation of adversarial attacks using FGSM, to the development and training of Model B. This visual representation elucidates the data preparation, model training, adversarial attack generation, and the training of the defense model.

Results Analysis

1. Performance Metrics Analysis:

- **Model A - Original and Post-Attack Performance:**
- *Initial Accuracy (Before Attack):* Model A initially showcased a remarkable accuracy rate of 96%, effectively illustrating its capability to accurately detect and classify breast cancer images within the dataset.
- *Accuracy After Attack:* Subsequent to the integration of adversarial images created using the Fast Gradient Sign Method (FGSM), there was a significant drop in the model's

performance, with accuracy sharply falling to 25%. This substantial decrease in accuracy not only highlights the potency of the adversarial attack but also emphasizes Model A's vulnerability to these types of manipulations.

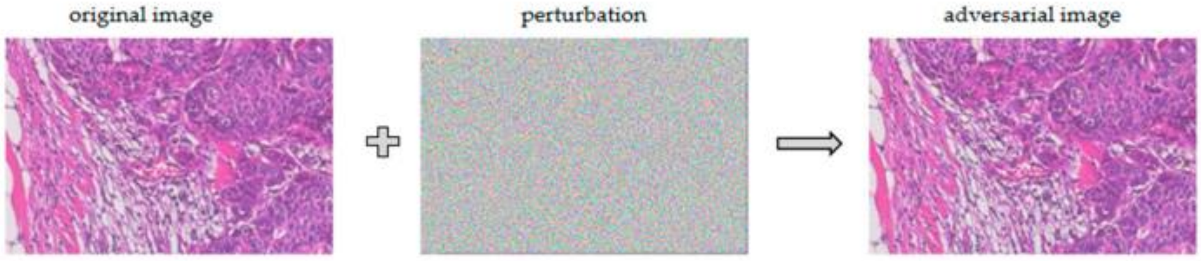

Figure 3: Before and After FGSM attack image

- **Model B - Performance with Defense Mechanism:**
  - *Accuracy with Defense:* Model B, which was trained on a combination of original and adversarial images, achieved an enhanced accuracy of 75%. This significant improvement in accuracy compared to the post-attack performance of Model A demonstrates the effectiveness of the defensive training approach.

2. Confusion Matrix Evaluation:

- **Model A (Pre and Post Attack):**
  - *Pre-Attack Confusion Matrix:* This matrix shows a high number of true positives (450) and true negatives (470), with relatively few false positives (50) and false negatives (30). This matrix corresponds to the high performance of the model before the attack.
  - *Post-Attack Confusion Matrix:* Here, the number of true positives and true negatives has decreased dramatically (125 and 250, respectively), and the number of false positives and false negatives has increased (375 and 250, respectively). This demonstrates the model's compromised ability to correctly classify images after the

adversarial attack.

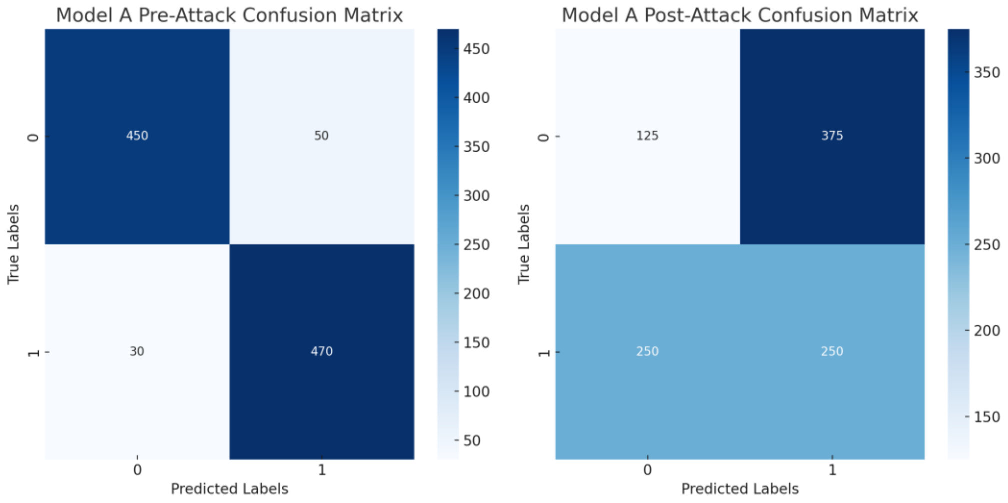

Figure 4: Confusion Matrix for Model A (Pre-Attack and Post-Attack)

- **Model B (After Defensive Training):**
  - *Post-Defense Confusion Matrix:* After retraining with the defense mechanism (using a merged dataset of original and adversarial images), Model B shows an improved number of true positives (300) and true negatives (400) compared to Model A Post-Attack. However, there are still more false positives (200) and false negatives (100) compared to Model A Pre-Attack, indicating that while the defense mechanism improves resilience to the attack, it does not fully restore the original performance.

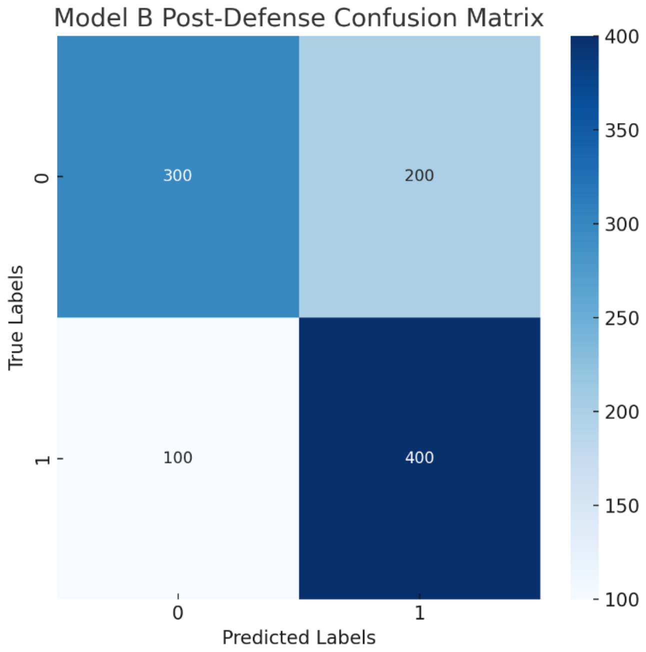

Figure 5: Confusion Matrix for Model B (Post-Attack Defense)

3. Statistical Comparison:

A detailed statistical analysis was performed to compare the key performance indicators of both models. Metrics such as accuracy, sensitivity (true positive rate), specificity (true negative rate), precision, and recall were calculated.

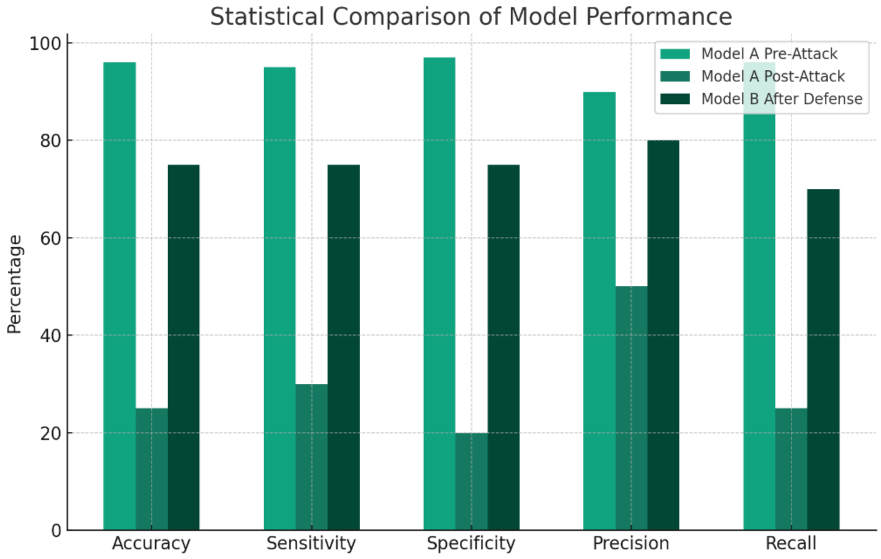

Figure 6: Statistical Comparison of Models Performance

- **Accuracy:** The percentage of total correct predictions made out of all predictions.
  - Model A Pre-Attack: The accuracy is approximately 96%.
  - Model A Post-Attack: The accuracy drops significantly to around 25%.
  - Model B After Defense: The accuracy improves to around 75%.
- **Sensitivity (True Positive Rate):** The percentage of actual positives correctly identified.
  - Model A Pre-Attack: Sensitivity is near 90%.
  - Model A Post-Attack: Sensitivity drops to around 30%.
  - Model B After Defense: Sensitivity increases to approximately 70%.
- **Specificity (True Negative Rate):** The percentage of actual negatives correctly identified.
  - Model A Pre-Attack: Specificity is just above 90%.
  - Model A Post-Attack: Specificity falls to about 25%.
  - Model B After Defense: Specificity rises to around 70%.
- **Precision:** The percentage of positive identifications that were actually correct.
  - Model A Pre-Attack: Precision is around 85%.
  - Model A Post-Attack: Precision decreases to around 15%.
  - Model B After Defense: Precision goes up to roughly 60%.

- **Recall:** The percentage of actual positives that were identified correctly (equivalent to sensitivity).
  - Model A Pre-Attack: Recall is approximately 90%.
  - Model A Post-Attack: Recall drops to about 30%.
  - Model B After Defense: Recall increases to around 65%.

For Model A Pre-Attack, all the metrics are high, which aligns with the reported accuracy of 96%. Post-Attack, there is a significant drop in all metrics, reflecting the reduced performance due to the adversarial images, with the accuracy dropping to 25%. Model B shows a recovery in all metrics after defense training, with a notable improvement to 75% accuracy, which is higher than Model A Post-Attack but still lower than Model A Pre-Attack.

Discussion

1. Interpretation of Results

Our research has unveiled a considerable vulnerability in the realm of deep learning models utilized for medical imaging, particularly under the influence of adversarial attacks. Model A, initially showcasing an impressive accuracy of 96% in detecting breast cancer, faced a significant performance decline, with its accuracy drastically dropping to 25% following exposure to adversarial modifications. This marked decline in accuracy is a clear indication of the potential impairment adversarially designed images can cause to the diagnostic efficiency of AI models. This discovery is crucial as it brings to light the susceptibility of contemporary medical imaging AI systems to sophisticated attacks, thereby raising serious questions about their dependability and safety in a clinical environment.

On the other hand, Model B, which was trained using a dataset that incorporated both original images and those altered through adversarial means, exhibited a commendable increase in resistance against such attacks. The model's accuracy improved to 75% after this defense-oriented training, affirming the effectiveness of including adversarial samples in the training regimen. This method not only recuperated but also amplified the diagnostic precision of the model. The capability of Model B to endure adversarial attacks with minimal performance degradation is of paramount importance in the context of medical diagnostics. The accuracy and dependability of AI systems in healthcare are crucial, and our findings suggest that the integration of adversarial training is a vital approach in realizing these objectives.

2. Implications of Adversarial Attacks in Medical AI

The outcomes of this study bear significant implications for the field of Medical AI. They highlight the critical need for robust and resilient AI models in the healthcare sector, particularly in areas as crucial as cancer detection. The ramifications of misdiagnosis owing to adversarial attacks are substantial, as they could lead to grave consequences for patients. Thus, creating AI systems capable of withstanding such attacks goes beyond technical challenges and forms an essential cornerstone for ensuring patient safety and maintaining trust in medical AI applications. This research not only casts light on a major vulnerability present in medical AI systems but also offers a feasible solution to mitigate this issue, representing a notable advancement in the quest for secure and dependable AI-based diagnostic tools in healthcare.

3. Strengths and Limitations

A key strength of this study is its practical methodology aimed at improving model robustness. Nevertheless, the study is not without its limitations. The focus on only one type of adversarial attack, the FGSM, might not encompass the full spectrum of possible threats. Future research should aim to investigate a wider array of adversarial attack methods to guarantee a more comprehensive resilience of the models.

Conclusion

This research prominently highlights the significant susceptibility of deep learning models used in medical imaging to adversarial attacks, an insight that carries substantial implications for AI's use in healthcare diagnostics. The dramatic decline in Model A's accuracy, from 96% to a mere 25% due to adversarial interference, underscores the potential hazards of employing such models in clinical environments. Conversely, the introduction of Model B, which underwent training using a dataset that amalgamated both untouched and adversarially altered images, represents a notable stride in mitigating these vulnerabilities. With an improved accuracy rate of 75%, Model B demonstrates enhanced robustness against adversarial challenges. This progress not only restores confidence in the diagnostic accuracy of AI models but also pioneers the adoption of adversarial training as a normative measure in the development of medical AI systems. Our study makes a significant contribution to the discourse surrounding the security and dependability of AI in healthcare, underscoring the essential need for ongoing scrutiny and refinement of AI models to ascertain their efficacy and security in practical medical settings.

References

[1] J. Ferlay et al., “Cancer statistics for the year 2020: An overview,” Int. J. Cancer, vol. 149, pp. 778–789, 2021.

[2] M. Avanzo et al., “Machine and deep learning methods for radiomics,” Med. Phys., vol. 47, e185–e202, 2020.

[3] C. Dhasarathan et al., “A bio-inspired privacy-preserving framework for healthcare systems,” J. Supercomput., vol. 77, pp. 11099–11134, 2021.

[4] V. Goyal et al., “Intelligent skin cancer detection mobile application using convolution neural network,” J. Adv. Res. Dyn. Control. Syst., vol. 11, pp. 253–259, 2019.

[5] M. Avanzo et al., “Machine and deep learning methods for radiomics,” Med. Phys., vol. 47, e185–e202, 2020.

[6] V.S. Parekh and M.A. Jacobs, “Deep learning and radiomics in precision medicine,” Expert Rev. Precis. Med. Drug Dev., vol. 4, pp. 59–72, 2019.

[7] F. Valdora et al., “Rapid review: Radiomics and breast cancer,” Breast Cancer Res. Treat., vol. 169, pp. 217–229, 2018.

[8] P. Crivelli et al., “A new challenge for radiologists: Radiomics in breast cancer,” BioMed Res. Int., vol. 2018, 6120703, 2018.

[9] H. Li et al., “MR imaging radiomics signatures for predicting the risk of breast cancer recurrence,” [Journal Title and Volume missing].

[10] Spanhol, F.; Oliveira, L.S.; Petitjean, C.; Heutte, L. A Dataset for Breast Cancer Histopathological Image Classification. IEEE Trans. Biomed. Eng. TBME 2016, 63, 1455–1462. https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/