Enhancing Breast Cancer Classification Robustness Through Adversarial Attack Training in Deep Learning Networks1

Kelly Chang

October 15, 2023

1 Background

October is Breast Cancer Awareness Month, which is one of the main reasons why I'm passionate about this topic. Breast cancer is a prevalent disease among women, and it ranks among the leading causes of cancer-related deaths in women. Early detection and treatment are crucial factors in improving survival rates. As a woman myself, I am committed to developing tools that aid in the diagnosis of breast cancer and facilitate prompt patient care.

In this project, I plan to train a deep learning neural network with adversarial attacks to classify breast ultrasound images. My goal is to increase the classification accuracy of the DNN and investigate the optimal level of perturbation required to enhance the robustness of breast cancer classification. Adversarial attacks typically involve making imperceptible changes to an image, yet these subtle alterations can lead the model to misclassify. By applying adversarial attacks to images and training DNN models with them, we can enhance the model's robustness. Previous research has successfully applied adversarial attacks to breast cancer histopathology images, resulting in improved accuracy [LL23].

I will be using the breast ultrasound image dataset I found on Kaggle. This dataset includes 780 images from 600 female patients aged between 25 and 75 years old. The images are categorized into three classes: normal, benign, and malignant. I will be using this label to generate a supervised classification model.

2 Model

Deep learning plays a crucial role in image classification tasks, and Convolutional Neural Networks (CNNs) stand out as one of the most widely used approaches. In my model, I integrated a simple CNN composed of three convolutional layers, each followed by ReLU activation functions and max-pooling layers, which helped in feature extraction and downsampling. The fully connected layers at the end

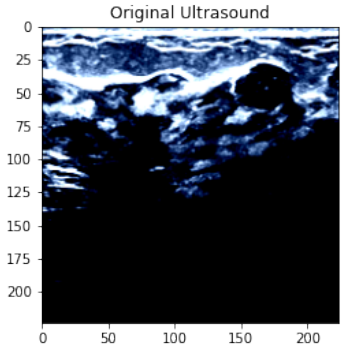

Figure 1: Ultrasound Image

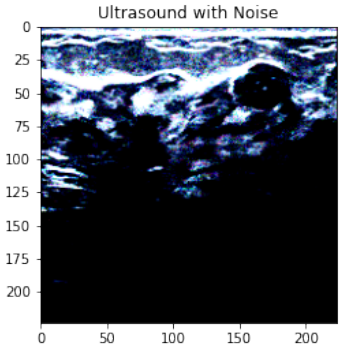

Figure 2: Ultrasound Image with Noise

processed the flattened feature maps to make predictions. Additionally, I harnessed the power of the built-in PyTorch ResNet50 to train on adversarial attacks, enhancing the robustness of the model.

In both models, I employed the cross-entropy loss function and the Adam optimizer for training. Furthermore, I trained each of them for 10 epochs with a batch size of 32 to enhance accuracy.

I utilized the train_test_split function from sklearn to partition the dataset into an 80% training set and a 20% testing set. Additionally, I crafted a function capable of structuring the training and testing sets into a format compatible with PyTorch's DataLoader.

I also implemented the Fast Gradient Sign Method (FGSM) to perturb my images. FGSM is a type of adversarial attack that leverages gradients to maximize the loss function and perturb the image. It operates as follows:

$$perturbed_x = x + \epsilon \cdot (\nabla_x Loss(\theta, x, y))$$

The reason why I choose FGSM is because it can generate adversarial examples quickly, which makes it suitable for testing and validating model robustness[NSKH23]. Following the implementation of this adversarial attack function, I generated new sets of perturbed training and testing data. Subsequently, I trained the models on the original training dataset and evaluated their performance on both the original testing dataset and the perturbed testing dataset.

3 Result

I chose $\epsilon = 0.5$ as the perturbation level for both the training and testing sets. The results, illustrated in Figures 3 and 4, demonstrate that training ResNet50 with the adversarially perturbed training set did not yield an enhancement in model accuracy. Instead, it led to an increase in accuracy variance (orange error bar).

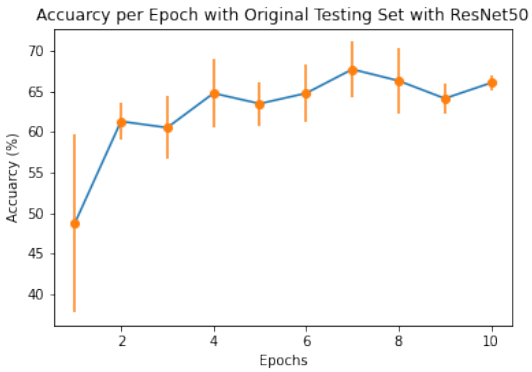

Figure 3: Accuracy of ResNet50

Figure 4: Accuracy of RestNet50 after Attack Training

Consequently, I initiated an investigation into the loss of ResNet50. Surprisingly, I found that the loss converged exceptionally well both before and after the adversarial training. Given that running ResNet50 is more computationally demanding than the simple CNN model, and it doesn't appear to excel at classification tasks, I decided to delve deeper into this issue using the simple CNN model.

However, when utilizing the simple CNN model, a significant boost in accuracy is evident when testing the adversarially trained model with the perturbed testing set, as illustrated in Figures 7 and 8. It is important to highlight that the accuracy of the original testing set decreased under these conditions. Notably, the loss also exhibited exceptional convergence both before and after the adversarial training.

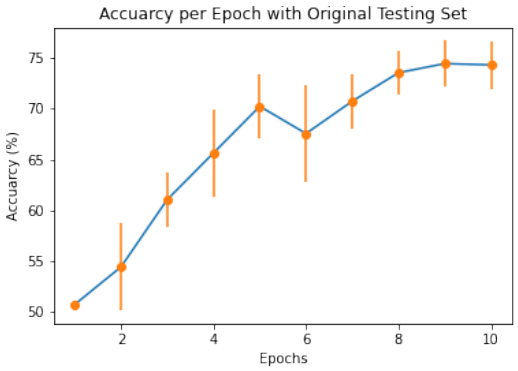

Figure 5: Accuracy of Simple CNN

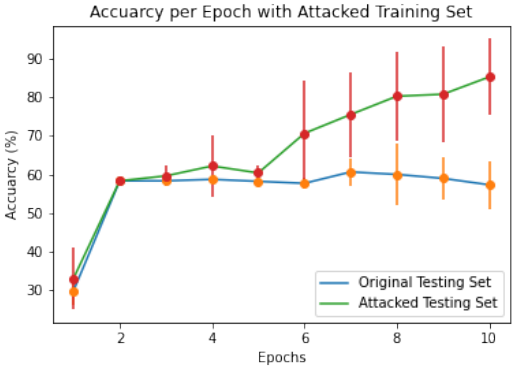

Figure 6: Accuracy of Simple CNN after Attack Training

Due to the adversarial trained model's difficulty in classifying unperturbed ultrasound images, I suspected that overfitting might be the issue. The model might be overly reliant on the added perturbations to classify each image, making it less effective at classifying original, unperturbed images.

To validate my hypothesis, I opted to retrain the model with smaller ϵ values, specifically $\epsilon = 0.2$ and $\epsilon = 0.05$. As illustrated in the figure below, when ϵ decreases, the adversarially trained model exhibits reduced accuracy in predicting perturbed images. However, it's important to note that this reduction in ϵ does not improve the model's ability to predict unperturbed images either. Therefore, while decreasing the ϵ value effectively addresses the overfitting issue associated with higher ϵ values, it does not contribute to enhancing the model's overall robustness.

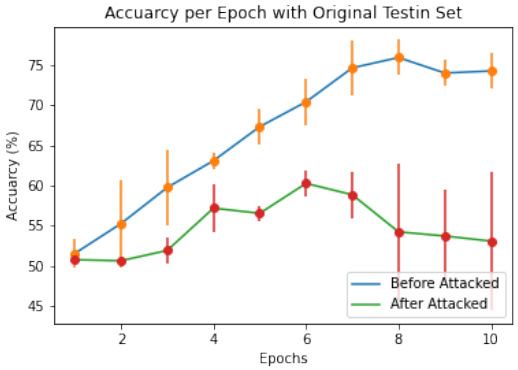

Figure 7: Accuracy on Test Set $\epsilon = 0.2$

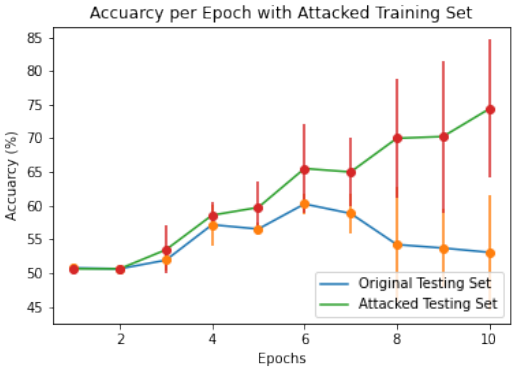

Figure 8: Accuracy after Attack Training $\epsilon = 0.2$

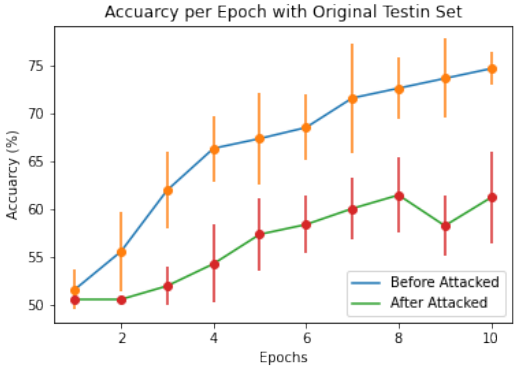

Figure 9: Accuracy on Test Set $\epsilon = 0.05$

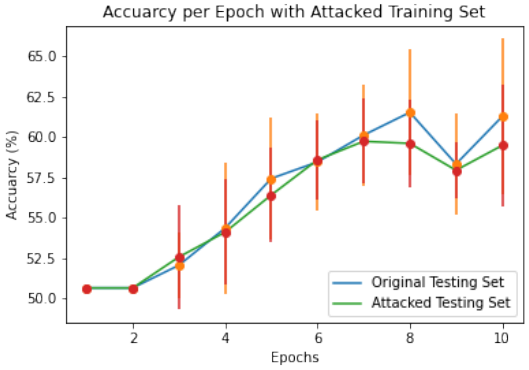

Figure 10: Accuracy after Attack Training $\epsilon = 0.05$

4 Conclusion

As demonstrated earlier, it is clear that adversarial attacks did not lead to an improvement in the robustness of my CNN model. This outcome could be attributed to various factors, such as the model's configuration, the possibility of an inadequate optimizer or loss function. To definitively determine the potential of adversarial attacks in enhancing model robustness, I envision conducting comprehensive testing involving a range of model settings in the future. Additionally, due to time constraints, I was unable to test ResNet50 with different values of ϵ. Training ResNet50 with varying ϵ values holds the potential to increase the robustness of breast cancer classification. This is another aspect I can explore in future research.

References

[LL23] Yang Li and Shaoying Liu. Adversarial attack and defense in breast cancer deep learning systems. *Bioengineering*, 10(8):973, 2023.

[NSKH23] Syed Muhammad Ali Naqvi, Mohammad Shabaz, Muhammad Attique Khan, and Syeda Iqra Hassan. Adversarial attacks on visual objects using the fast gradient sign method. *Journal of Grid Computing*, 21(4), 2023.