Red Teaming analysis of the Breast Cancer Detector Model

Mikołaj Drzewiecki md406134@students.mimuw.edu.pl
Monika Michaluk mm395135@students.mimuw.edu.pl
University of Warsaw, Poland

Abstract

This study conducts a Red Teaming analysis on the Breast Cancer Detector Model, a Convolutional Neural Network designed for predicting breast cancer from tissue scans. Using eXplainable Artificial Intelligence (XAI) techniques, we assess the model's reliability, investigating influences from unintended artifacts and evaluating its generalization with out-of-distribution samples. Our aim is to uncover vulnerabilities and enhance the model's robustness in clinical applications.

1. Introduction

In recent years, Red Teaming has emerged as a crucial methodology for evaluating the robustness and reliability of machine learning models. With the widespread adoption of deep learning models, a comprehensive understanding of their decision-making processes is important. Especially, this becomes crucial in medical applications, where artifacts or biases in the training data can lead to inaccurate predictions, posing significant risks to patient outcomes. Additionally, a clear insight into a model's predictions is important for clinicians to confidently integrate AI-based tools into their decision-making processes.

Our analysis focuses on a comprehensive examination of a Breast Cancer Detector Model. To test how well the model works, we examined challenging samples and conducted out-of-distribution testing by augmenting images from the training set with variations in brightness and color. We used LIME and SHAP to make sure the model's predictions were based on real cancer-related features. By comparing with examples from a dataset with annotations, we could understand more about how the model makes decisions. Through these methods, we found out important features and weaknesses in the model, helping us understand better where it works well and where it might struggle in medical diagnoses.

2. Methodology

2.1 Model

We examine a classification model based on Convolutional Neural Networks (CNNs) available on Hugging Face. The model is designed to predict whether breast tissues within image patches contain Invasive Ductal Carcinoma (IDC) - the predominant type of breast cancer. The model attains an accuracy of 87% on the test set.

2.2 Datasets

We examine the model using two datasets:

Breast Histopathology Images: The training dataset, available on Kaggle. It consists of 277,524 patches sized 50 x 50 extracted from 162 whole mount slide images of Breast Cancer specimens scanned at a magnification of 40x. This patches are categorized into IDC negative (198,738 patches) and IDC positive (78,786 patches).

BreCaHAD: The original dataset, from which the patches were extracted, comprises 162 breast cancer histopathology microscopic biopsy images with dimensions of 1360x1024 pixels. Each image is annotated, a crucial aspect for comparison with generated explanations.

2.3 Data Augmentation and Dataset Exploration

We conduct a basic investigation of the model's performance on the training dataset, including a review of the confusion matrix and logit distribution. We investigate positive and negative samples visually to identify patterns, easy and hard samples or samples for which the model is not certain of it's prediction. We generate histograms representing the distribution of pixel values for positive and negative predictions and compare them with the true distribution. We apply common data perturbation techniques, such as Gaussian blur and variations in color and brightness, to the original dataset, and evaluate the model's performance on such modified images. Additionally, we test the model on a subset of the BreCaHAD dataset on randomly cropped 50x50 patches.

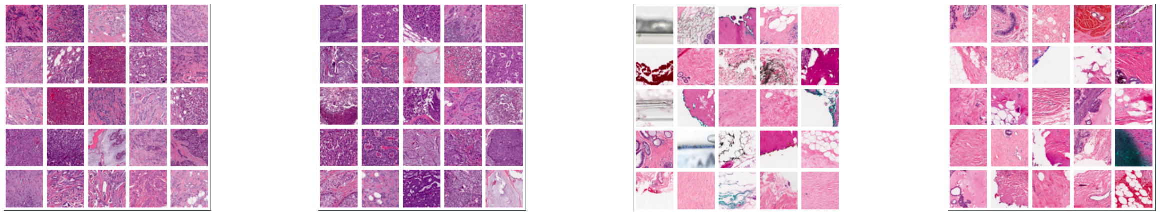
2.4 Explanation techniques

LIME is a technique for generating locally interpretable explanations of machine learning models. It works by perturbing input instances and observing changes in predictions. In our study, LIME is applied to specific dataset samples, creating simple, understandable models for each instance. This approach helps reveal the factors influencing the model's decision-making at the local level.
SHAP values, rooted in cooperative game theory, allocate the contribution of each feature fairly to a prediction. In model interpretation, SHAP values quantify the impact of individual features on the model's output. For our analysis, SHAP is used to understand the importance of different features in the Breast Cancer Detector Model and highlight critical elements influencing the model's decision.

3. Experimental results

3.1 Initial investigation

Initially, we confirmed the model's accuracy at 87% on the provided dataset. We examined the confusion matrix of the predictions: total number of true negatives (1911196), false negatives (27941), true positives (7542) and true positives (50845). Notably, there is a relatively high amount of false negatives - approximately one-third of cancerous tissues are predicted as healthy. Further analysis of the distribution of logits showed the model's



(a) TP (b) FP (c) TN (d) FN

Figure 1: Comparison of positive and negative samples

high certainty in negative predictions but uncertainty in positive predictions, potentially attributed to the initial class imbalance in the training dataset.

Further, we visually examined samples where the model made incorrect predictions, revealing a pattern in image style. Specifically, samples predicted as negative consistently exhibited lighter colors and a smooth structure, while samples predicted as positive tended to be darker and coarser. Uncertain samples were a combination of the two. Example samples are showed on Figure 1.

To test our hypothesis, we conducted a simple experiment with 1000 random images. We estimated pixel value distributions for samples labeled as positives and compared them with the distributions for samples predicted as positive by the model. The same comparison was made for negative samples, revealing strikingly similar distributions. We include distribution plots in Appendix A.

3.2 Data augmentation

We assessed the model's robustness to dataset perturbations, including grayscale conversion, Gaussian Blur application, and brightness adjustment. The evaluation involved 1000 randomly selected samples, revealing the model's high sensitivity to these perturbations. Grayscale conversion consistently led to false labels, while adjusting brightness resulted in a significant accuracy drop. Beyond a certain threshold, the model consistently assigned false labels again. Interestingly, Gaussian Blur didn't show this effect - the model remained quite robust to blurred images. Plots illustrating accuracy across different values of blur/brightness factor can be found in Appendix B.

3.3 LIME

Analysis performed with LIME revealed that the model assigns the greatest contribution to regions lying between light and dark areas. Those parts of the image contribute to the overall prediction the most. This tendency is independent to image's brightness, color palette etc. We observed it while explaining both correct and inaccurate predictions - of both healthy and infected tissues. The model tends to associate images, which depict sharp transitions, with high probability of cancer. Although the huge majority of test images indeed could be classified this way, we were able to fool the model, providing images (of infected tissues) with "mild" transitions, described above. The result usually lay somewhere in range (0.1 - 0.4), while to be classified as infected tissue - the result is required to be above 0.5.

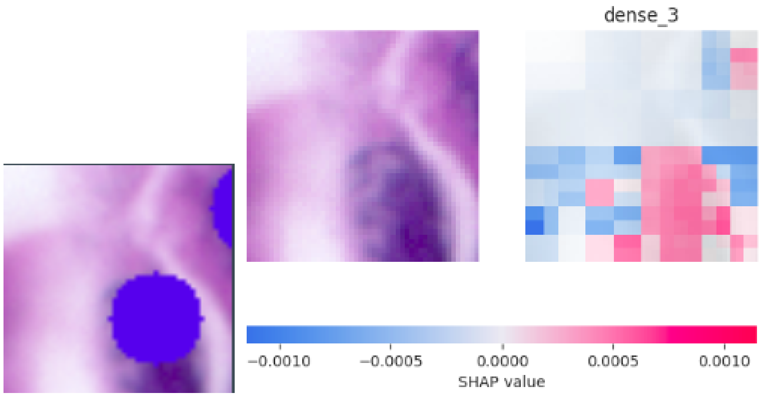


Figure 2: Comparison of SHAP feature attribution with annotated sample

3.4 SHAP

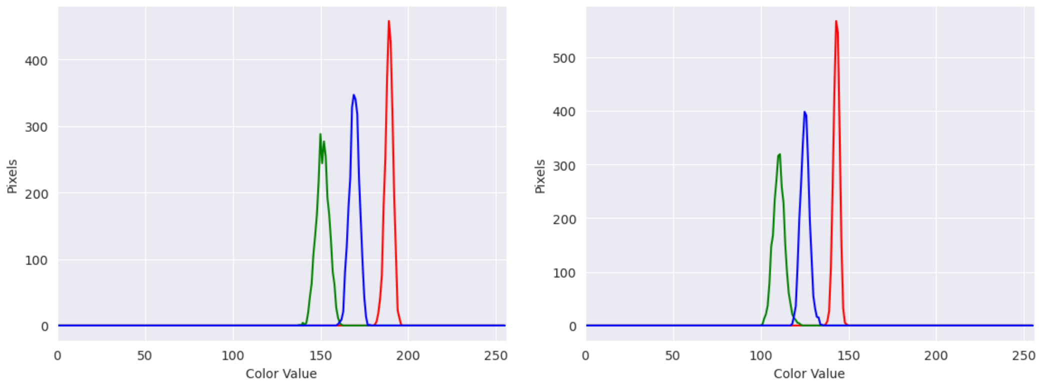
We tried to find out, how little changes of image's brightness can affect the contribution of same areas to the prediction. Regardless of the brightness level (within the range within which the model achieves > 80% accuracy), SHAP analysis showed that the model is almost always able to spot distinctive areas and correctly associate them with positive/negative influence. We investigated the arising question - how changes in brightness affect largeness of the SHAP values for every spotted area. Again, the model proved to behave almost identically. The SHAP values we intended to test, ranged usually from -0.001 to 0.001 for every pixel. We also observed the phenomenon discovered with LIME - pixels belonging to distinctive perimeters or hulls affected the result the most.

We observed, however, if not vulnerability - an interesting property of bright areas. It turns out that alongside with the rising prediction value, the model tends to associate increasingly significant contribution to distinctively lighter areas (especially when they are very small). This behavior resembles the mirrored problem - as if the model was to detect healthy region in cancer tissue. With no doubt intriguing, the influence of this property proved to be exceptionally elusive, affecting the prediction little to nothing.

4. Conclusion

Performed analysis proved strength of the model. Various vulnerabilities we were able to find turned out to be very intricate, unlike to be revealed in day to day usage. We assess the uncertainty while classifying the sample as infected tissue, to be promising field to improve. We can conclude that phenomena discovered during LIME and SHAP analysis are deeply connected to this problem. Also, we advise care in ensuring the quality of input data, as vulnerabilities involving data augmentation were exposed.

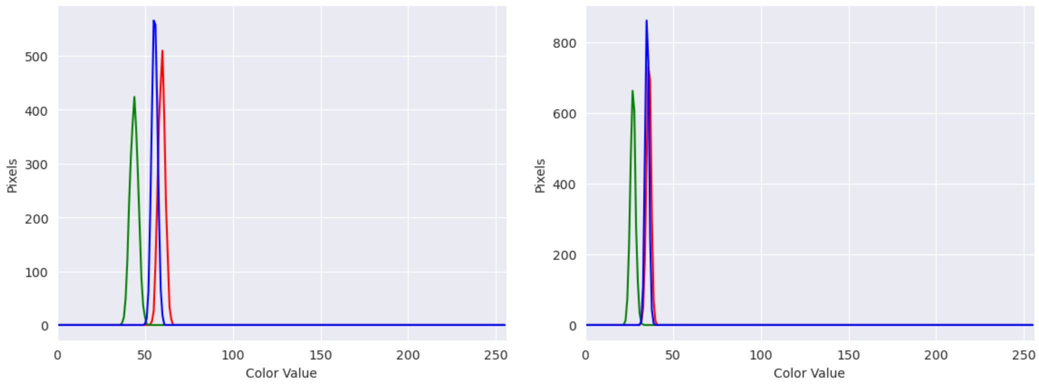
References



(a) True label: 0

(b) Prediction: 0

Figure 3: Comparison of color distribution in ”no-cancer” samples



(a) True label: 1

(b) Prediction: 1

Figure 4: Comparison of color distribution in ”cancer” samples

Appendix A. Estimated color distribution

Appendix B. Performance on perturbed data

Appendix C. SHAP plots

todo

Appendix D. LIME plots

todo