Detecting Adversarial Attacks On Breast Cancer Diagnostic Systems Using Attribution-based Confidence Metric

Steven L. Fernandes
Computer Science Department
Creighton University, NE
stevenfernandes@creighton.edu

Senka Krivic
Faculty of Electrical Engineering
University of Sarajevo, B&H
senka.krivic@etf.unsa.ba

Poonam Sharma
Pathology Department
Creighton University, NE
poonamsharma@creighton.edu

Colleen Westerhaus
Business Intelligence and Analytics
Creighton University, NE
colleenwesterhaus@creighton.edu

Sumit K. Jha
Computer Science Department
University of Texas at San Antonio, TX
sumit.jha@utsa.edu

Abstract

In this paper, we develop attribution-based confidence (ABC) metric to detect black-box adversarial attacks in breast histopathology images. Due to the lack of data for this problem, we subjected histopathological images to adversarial attacks using the state-of-the-art technique Meta-Learning the Search Distribution (Meta-RS) and generated a new dataset. We adopt the Sobol Attribution Method to the problem of cancer detection. The output helps the user to understand those parts of the images that determine the output of a classification model. The ABC metric characterizes whether the output of a deep learning network can be trusted. We can accurately identify whether an image is adversarial or original with the proposed approach validating it with eight different deep learning-based classifiers. The ABC metric for all original images is greater or equal to 0.8 and less for adversarial images. To the best of our knowledge, this is the first work to detect attacks on medical systems for breast cancer detection using the ABC metric.

1 Introduction

Breast cancer is the most common type in women, with 1.68 million registered modern cases and 522,000 caused deaths in 2012 [1–3]. Histopathological image analysis systems provide precise models and accurate quantification of the tissue structure [4]. To provide automatic aid for pathologists, deep learning networks are used for tracing cancer signs within breast histopathology images [3, 5]. Generative Adversarial Networks (GANs) are used to generate new digital pathology images [6]. However, this brings a high risk of medical image analysis systems being subject to black-box adversarial attacks. Adversarial images are hard to detect and can easily trick human users and AI systems. Solving it leads to more secure medical systems and more explainable systems [7]. We exploit Sobol Attribution Method for explanations [8] which captures interactions between image regions and is used to visualize how they affect the neural network's prediction. Due to the specificity of the pathological images, additional information is needed to detect adversarial attacks. We develop attribution-based confidence (ABC) measure [9] to quantize the decision of whether an image is original or not. We perform an adversarial attack using the state-of-the-art method, Meta-Learning the Search Distribution (Meta-RS) of Black-Box Random Search [10] on images from BreaKHisdatabase [11].

[0]The new dataset is available at https://bit.ly/3p4QaPw

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

2 Proposed Approach

Firstly, we trained eight architectures for the image classification task (ResNet18, ResNet50, Inception V3, MobileNet V3, ShuffleNet, Swin Transformer, Vision Transformer, WideResnet) with the original dataset. In the second step, we performed state-of-the-art Meta-RS black-box adversarial attacks [10] and generated an adversarial images dataset. Then, we adapted the Sobol Attribution Method for explanations [8]. Finally, we proposed attribution-based confidence (ABC) metric [9] to detect black-box adversarial attacks, (Figure 1).

Adversarial Images Generation: We pose the problem as a meta-learning problem following the work by Yatsura et al. [10]. For the dataset $(x, y) \in \mathcal{D}$, classifier models $f \sim \mathcal{F}$, and the stochastic adversarial perturbation ϵ^{ω} the meta-objective is to determine parameters ω^* of the attack $\mathcal{A}_{\omega}$ [10]. This learning approach is applied to Square Attack [12] with l_{∞} threat and is called Meta Square Attack (MSA). MSA operates with computation of the square size in pixels with size controllers $s_t = \pi^s_{\omega_s} \in \{1, ..., s_{max}\}$ and sampling position $(p_x, p_y) \sim \pi^p(s) \in \{1, ..., s_{max} - s\}^2$ and sampling color with a color controller $c \sim \pi^c_{\omega_c} \in \{c_1, ..., c_m\}$. Position controller π^p is uniform distribution while color and size controllers are meta-learned multi-layer perceptron (MLP) with parameters ω_s and ω_c.

Sobol Attribution Method describes the decision of a black-box system $f : \mathcal{X} \rightarrow \mathcal{R}^k$ based on the given input image described with a collection of features $x = (x_1, ..., x_n)$. The method exploits the random perturbations approach to determine interactions among the features and their contribution to $f(x)$. The perturbations are defined with a probability space $\Omega, \mathcal{X}, P$ of possible input perturbations and a random vector $X = (X_1, ..., Xn)$ on the data manifold around the input x. For their values holds $\sum_{u \in \mathcal{U}} \mathcal{S}_u = 1$. The total Sobol index $\mathcal{S}_{T_i}$ defines how the variable X_i affects the model output variance and the interactions of any order of X_i with any other input variables.

Attribution Based Confidence (ABC) Metric characterizes whether one can trust the decision of a deep neural network on an input [9]. The concentration of features characterizes DL models. The assumption is that sampling over low features will result in no change in the output. Low attribution provides information that the model is *equivariant* along the features. For an input x, a classifier model f,we compute attribution of the feature x_j of x as $A_j(x)$. The ABC metric is calculated by: *(i)* sampling the neighbourhood and *(ii)* measuring the conformance. Sampling is done by selecting the vector x_j with the probability of $P(x_j)$ and changing its value can result in a change in the model's decision. The procedure is repeated S times for the input image.

Algorithm 1 Generate Adversarial Images

Input: Data distribution $\mathcal{D}$, a classifier f, number of epochs, SA budget, uniform distribution π^p, **Output:** generated images $\mathcal{D}_g$

1: **for** number of epochs **do**
2: $\quad \pi^s_{\omega_s} \leftarrow trainMLP(D, SAbudget)$
3: $\quad \pi^c_{\omega_c} \leftarrow trainMLP(D, SAbudget)$
4: **end for**
5: **for** number of attacks **do** $\mathcal{D}_g \leftarrow \mathcal{D}_g \cup SA(\pi^p, \pi^s_{\omega_s}, \pi^c_{\omega_c})$
6: **end for**

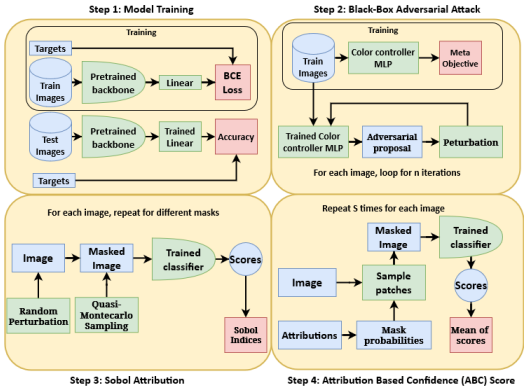

Figure 1: Proposed approach for detecting adversarial images

Algorithm 2 Total Order Estimator (Pythonic implementation)

Input: Prediction scores Y, dimension $d = 8 \times 8$, number of designs N
Output: Total Sobol Index S_{T_i}

1: $f(A) = Y[1 : N]$, $f(B) = Y[N : N * 2]$ (perturbed inputs)
2: **for** i=1 to d **do**
3: $\quad f(C) = Y[N * 2 + N * i : N * 2 + N * (i + 1)]$
4: **end for**
5: $f_0 = \frac{1}{N} \sum_{j=0}^{N} f(A_j)$
6: $\hat{V} = \frac{1}{N-1} \sum_{j=0}^{N} (f(A_j) - f_0)^2$
7: $S_{T_i} = \frac{\frac{1}{2N} \sum_{j=0}^{N} (f(A_j) - f(C_j^{(i)}))^2}{\hat{V}}$

Algorithm 3 Calculate ABC Metric

Input: a classifier f, input x, sample size S
Output: ABC metric $c(f, x)$

1: $A_1, ..., A_n \leftarrow$Attributions of $x_1, ..., x_n$ from x
2: $i \leftarrow f(x)$ (get classification output)
3: **for** j = 1 to n **do**
4: $\quad P(x_j) \leftarrow \frac{|A_j / x_j|}{\sum_{k=1}^{n} |A_k / x_k|}$
5: **end for**
6: Generate S from mutation of x_j with $P(x_j)$
7: Get classification output for S samples
8: $c(f, x) \leftarrow S_{conform}/S$

3 Experimental results

We have selected 1148 microscopic images from the Breast Cancer Histopathological Image Classification (BreakHis), the dataset of breast tumor tissue images collected from 82 patients using a 400x magnifying factor. Train and test datasets correspond to 80% and 20% of the data, respectively. All correctly predicted samples were under attack during the testing. All correctly classified samples have been modified for 1000 iterations for each case.

	Model	Train accuracy	Total attacks	Successful attacks	Attack accuracy
Train	InceptionV3	0.917	841	144	0.171
	ResNet18	0.968	888	270	0.304
	ResNet50	0.999	915	565	0.617
	MobileNet V3	0.939	861	280	0.325
	ShuffleNet	0.995	913	273	0.299
	Swin Transformer	0.985	904	218	0.241
	Vision Transformer	0.999	917	236	0.257
	Wide ResNet	0.992	908	278	0.306
Test	InceptionV3	0.848	195	51	0.261
	ResNet18	0.882	202	56	0.276
	ResNet50	0.900	208	49	0.236
	MobileNet V3	0.839	193	54	0.280
	ShuffleNet	0.904	208	56	0.269
	Swin Transformer	0.935	215	49	0.228
	Vision Transformer	0.891	205	40	0.195
	Wide ResNet	0.904	208	58	0.279

Table 1: Classification accuracy for the train and test datasets subjected to state-of-the-art Meta-RS black-box adversarial attacks.

Sobol Attribution Method was used on the images with masks generated with resolution $d = 8 \times 8$. The same resolution was used occlusion to sign the $\hat{S}_{T_i}$. Zero was used for the baseline. The number of designs was set to $N = 32$. As perturbation function was used Inpainting. ABC metric[1] parameter of the sample size was set to $S = 1000$. We performed classification tests to evaluate the success of adversarial attacks done with the Meta-RS algorithm for previously correctly classified samples. The attack results for the training dataset and for the test dataset are in Table 1. Explanations generated with Sobol Attribution Method represent a visual aid for a user to understand which regions of the images affected the decision-making. In Fig. 2 for successful attacks in the case of ResNet50 and corresponding original images. Table 2 provides ABC metric values for the original and adversarial images for all eight classification models. All experiments were conducted on Google Colab Pro+ with NVIDIA T4 Tensor Core GPU and 52 GB RAM.

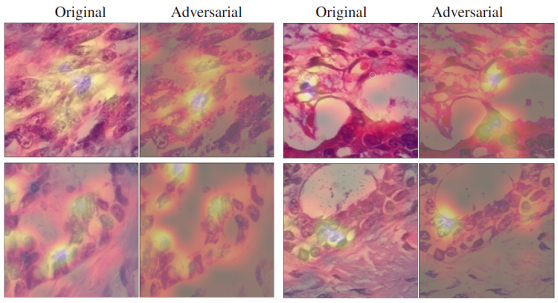

Figure 2: Sobol attribution explanations for ResNet50.

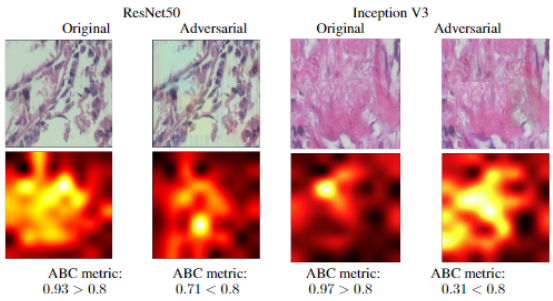

Figure 3: ABC values are used to differentiate between original and adversarial images.

4 Conclusions

We subjected classification models to state-of-the-art robust Meta-RS attacks. The obtained adversarial images are available for public use. Sobol Attribution Method [8] was applied to understand those parts of the images that determine the output of a classification model. However, due to the nature of histopathological images and the specificity of the classification problem, several regions are being highlighted with the Sobol attribution method.

	Attribution-based Confidence (ABC)			
	Train dataset		Test dataset	
Model	Original	Adversarial	Original	Adversarial
ResNet18	0.920	0.539	0.948	0.518
ResNet50	0.928	0.340	0.908	0.323
Inception V3	0.847	0.734	0.893	0.698
MobileNet	0.876	0.357	0.861	0.389
ShuffleNet	0.934	0.742	0.930	0.732
Swin Transformer	0.971	0.702	0.969	0.710
Vision Transformer	0.947	0.731	0.945	0.722
Wide ResNet	0.893	0.034	0.874	0.013

Table 2: ABC metric values for the eight models

To the best of our knowledge, this is the first work to detect attacks on medical systems for breast cancer prediction based on histopathological images using the ABC metric. The evaluation of eight different classification models shows that the ABC metric for all original images is greater or equal to 0.8 and less than 0.8 for adversarial images.

[1] https://github.com/ma3oun/abc_metric

Potential Negative Societal Impact

We do not foresee many potential negative societal impacts to our work. Given the importance of robust and reliable predictions in the medical imaging domain, we believe detecting adversarial attacks is important for any model that is to be deployed in practice.

References

[1] Z. Senousy, M. M. Abdelsamea, M. M. Gaber, M. Abdar, U. R. Acharya, A. Khosravi, and S. Nahavandi, "Mcua: Multi-level context and uncertainty aware dynamic deep ensemble for breast cancer histology image classification," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 2, pp. 818–829, 2022.

[2] P. Thiagarajan, P. Khairnar, and S. Ghosh, "Explanation and use of uncertainty quantified by bayesian neural network classifiers for breast histopathology images," *IEEE Transactions on Medical Imaging*, vol. 41, no. 4, pp. 815–825, 2022.

[3] H. Yang, J.-Y. Kim, H. Kim, and S. P. Adhikari, "Guided soft attention network for classification of breast cancer histopathology images," *IEEE Transactions on Medical Imaging*, vol. 39, no. 5, pp. 1306–1315, 2020.

[4] C. Mercan, B. Aygunes, S. Aksoy, E. Mercan, L. G. Shapiro, D. L. Weaver, and J. G. Elmore, "Deep feature representations for variable-sized regions of interest in breast histopathology," *IEEE Journal of Biomedical and Health Informatics*, 2021.

[5] M. Liu, L. Hu, Y. Tang, C. Wang, Y. He, C. Zeng, K. Lin, Z. He, and W. Huo, "A deep learning method for breast cancer classification in the pathology images," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–8, 2022.

[6] A. Das, V. K. Devarampati, and M. S. Nair, "Nas-sgan: A semi-supervised generative adversarial network model for atypia scoring of breast cancer histopathological images," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 5, pp. 2276–2287, 2022.

[7] X. Ma, Y. Niu, L. Gu, Y. Wang, Y. Zhao, J. Bailey, and F. Lu, "Understanding adversarial attacks on deep learning based medical image analysis systems," *Pattern Recognit.*, vol. 110, p. 107332, 2021.

[8] T. Fel, R. Cadène, M. Chalvidal, M. Cord, D. Vigouroux, and T. Serre, "Look at the variance! efficient black-box explanations with sobol-based sensitivity analysis," in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021, pp. 26 005–26 014.

[9] S. Jha, S. Raj, S. L. Fernandes, S. K. Jha, S. Jha, B. Jalaian, G. Verma, and A. Swami, "Attribution-based confidence metric for deep neural networks," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2019, pp. 11 826–11 837.

[10] M. Yatsura, J. Metzen, and M. Hein, "Meta-learning the search distribution of black-box random search based adversarial attacks," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 30 181–30 195.

[11] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1455–1462, 2016.

[12] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, "Square attack: A query-efficient black-box adversarial attack via random search," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIII*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12368. Springer, 2020, pp. 484–501.