Attribution-Based Confidence Metric for Detection of Adversarial Attacks on Breast Histopathological Images

Steven L. Fernandes[1], Senka Krivic[2], Poonam Sharma[3], and Sumit K. Jha[4]

[1] Computer Science Department, Creighton University, NE
stevenfernandes@creighton.edu
[2] Faculty of Electrical Engineering, University of Sarajevo, B H
senka.krivic@etf.unsa.ba
[3] Pathology Department, Creighton University, NE
poonamsharma@creighton.edu
[4] 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. The proposed approach is validated 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 based on histopathological images using the ABC metric.

Keywords: deep learning, adversarial attacks, explainable AI

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 [18,20,23]. Histopathological image analysis systems provide precise models and accurate quantification of the tissue structure [16]. To provide automatic aid for pathologists, deep learning networks are used for tracing cancer signs within breast histopathology images [23,13]. Moreover, Generative Adversarial Networks (GANs) are used to generate new digital pathology images [4]. 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. Therefore, detecting adversarial images and any artificial change to the original image is a crucial problem in medical image analysis. Solving it leads to more secure medical systems and more explainable systems [14].

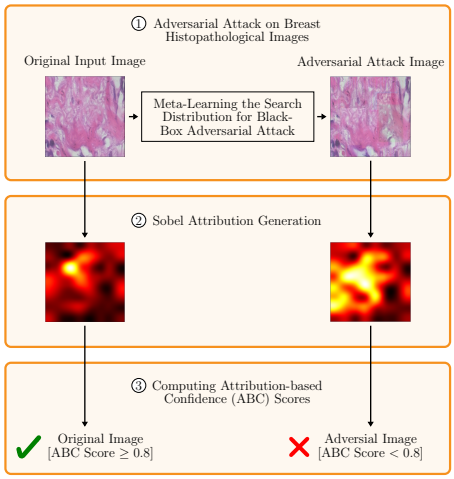

Fig. 1. The adversarial attack detection process.

We exploit Sobol Attribution Method for explanations [5] which captures interactions between image regions with Sobol indices 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 [11] to quantize the decision of whether an image is original or not (Fig. 1). We demonstrate how to perform an adversarial attack using the state-of-the-art method, Meta-Learning the Search Distribution (Meta-RS) of Black-Box Random Search [24] on digital pathology images from BreaKHis [5] database [19]. Using transfer learning, we evaluated the proposed approach with eight pre-trained classifiers on both the train and test datasets.

The main contributions of this paper are:

1. We created an adversarial histopathological image dataset using the state-of-the-art black-box attacks, and we make it public for further use [6].
2. We use the state-of-the-art Sobel attribution-based method to understand those parts of the images that determine the output of a classification model.
3. We developed the ABC metric, which indicates original images when the metric values are greater than or equal to 0.8 and adversarial images otherwise. This validated eight deep learning-based classification models and trained and tested data.

To the best of our knowledge, this is the first work to detect attacks on medical systems for breast cancer detection based on histopathological images using an attribution-based confidence metric (ABC) metric.

[5] https://www.kaggle.com/datasets/forderation/breakhis-400x
[6] The new dataset is available at https://bit.ly/3p4QaPw

2 Related Work

Histopathology image analysis plays a critical role in cancer diagnosis and treatment. The development of deep neural networks has made many breakthroughs in challenging clinical tasks such as automatic image analysis and prediction and assisted diagnosis [22,26,8,15].

Xu et al. [22] proposed a weakly supervised learning framework for segmentation of histopathology images using just image-level labels. The results demonstrate remarkable performance with the fully supervised approaches. Most recent approaches combine several methods to improve the performance and capability of automatic diagnosis.

In this way, Hashimoto et al. [8] proposed a new CNN-based approach which can classify cancer sub-type for cancer sub-type from histopathology images. With the proposed method, it is possible to examine the whole slide and, in an automated way, detect tumour-specific features. Their method combines domain adversarial, multiple-instance, and multi-scale learning frameworks.

Zhao et al. [26] address the problem of automatic lymph node metastasis prediction. histopathological images of colorectal cancer. They created a GCN-based Multiple Instance Learning methods with a feature selection strategy. Lack of image datasets and cost of image annotation is often a limitation in the medical imaging field [9,21]. Therefore there is a large number of works dedicated to decreasing this problem either with proposed approaches for image generation, automated image labelling or segmentation [9,1].

Gamper and Rajpoot [1] present a novel multiple instances captioning dataset to facilitate dense supervision of CP tasks. This dataset contains diagnostic and morphological descriptions for various stains, tissue types and pathologies. Experimental results demonstrate that their proposed representation transfers to a variety of pathology sub-tasks better than ImageNet features or representations obtained with learning on pathology images alone.

Other methods for improving the automated pathology decision-making process include also automatizing preprocessing steps such as automatic magnification selection [25]. With the increase of the automatizing pathological diagnosis and application of AI methods, data and results malversation is potentially risky. Adversarial attacks are considered a potentially serious security threat for machine learning systems [3,14].

Fote et al. [6] show that a highly accurate model for classifying tumour patches in pathology images can easily be attacked with minimal perturbations which are imperceptible to lay humans and trained pathologists alike. Therefore, there is a need to detect adversarial attacks and increase the security of medical systems. Laleh et al [12] show that CNNs are susceptible to various white- and black-box attacks in clinical cancer detection tasks. Paschali et al. [17] demonstrate that besides classification tasks, also segmentation tasks can be affected by adversarial attacks. Thus, they propose a model evaluation strategy by leveraging task-specific adversarial attacks.

After reviewing the existing literature, we found a lack of adversarial breast histopathological image datasets and robust techniques to detect such attacks.

3 Proposed Approach

This section details methods for developing the approach for adversarial attack detection in breast histopathology images. Figure 2 provides a detailed overview of the process of developing the approach for adversarial image detection.

In the first step, 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 [24] and generated an adversarial images dataset. In the third step we adapted the state-of-the-art Sobol Attribution Method for explanations [5].

Finally, we propose attribution-based confidence (ABC) metric [11] to detect black-box adversarial attacks. The ABC metric characterizes whether one can trust the decision of a deep neural network on an input.

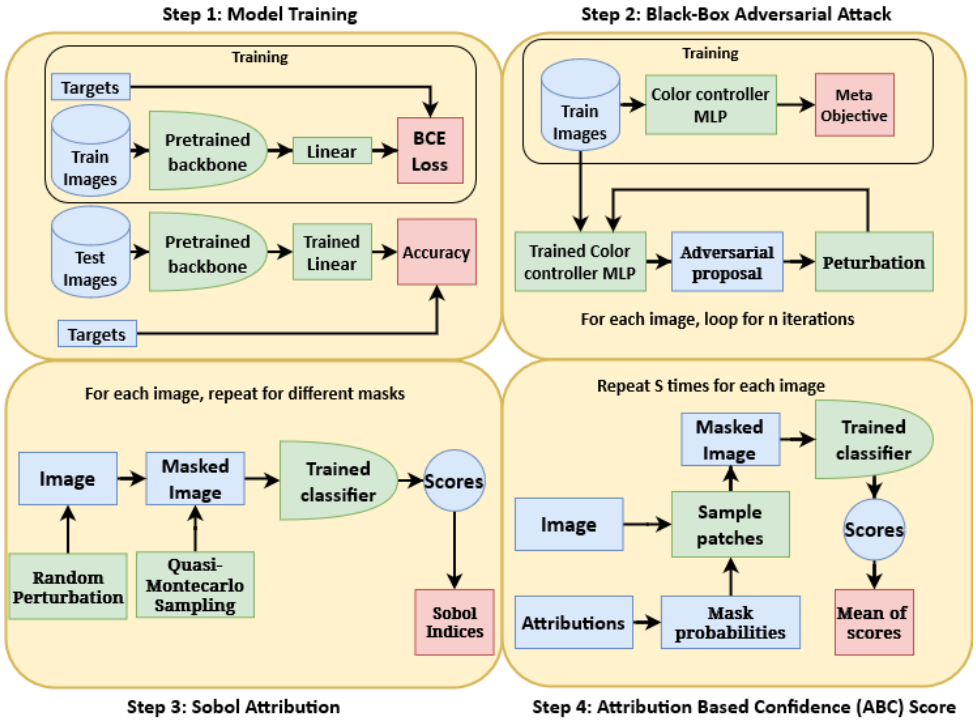

Fig. 2. Overview of the proposed approach for detecting adversarial images

Adversarial Images Generation is commonly done with Generative Adversarial Networks (GANs). We pose the problem of adversarial image generation as a meta-learning problem following the work by Yatsura et al. [24]. 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 $\mathcal{A}_{\omega}$.

Determination of parameters of the attack is done through maximisation of the lower bound $L(f, x, y, \epsilon^{\omega})$ of the goal function $V(f, x, y)$.

$$\omega^* = \underset{\omega}{\operatorname{argmax}} \mathop{\mathbb{E}}_{f \sim \mathcal{F}(x,y)} \mathop{\mathbb{E}}_{\sim \mathcal{D} \epsilon^{\omega} \sim \mathcal{A}_{\omega}(L,f,x,y)} L(f, x, y, \epsilon^{\omega}) \tag{1}$$

The meta-representation is defined such that A_{ω} effectively generalizes across models $f \sim \mathcal{F}$. For a random search-based attack where the query budget is determined by a limit T, an adversarial perturbation on the perturbation set S is defined with an iterative process:

$$\epsilon^0 \sim \mathcal{D}^0; \epsilon^0 t + 1 = \underset{\epsilon \in \{\epsilon^t, \mathcal{P}_s(\epsilon^t + \delta^{t+1})\}}{\operatorname{argmax}} L(f, x, y, \epsilon); \delta^{t+1} \sim \mathcal{D}_{\omega}(t, \epsilon^0, \delta^0, ..., \epsilon^t, \delta^t) \tag{2}$$

where $\mathcal{P}_{\mathcal{S}}$ is a projection on the perturbation set S.

With an assumption that loss function l and A_{ω} are differentiable with respect to the meta-parameters ω, the meta-optimization for determining meta-parameters is done with stochastic gradient descent optimization on batches $B \subseteq D$ based on the gradient:

$$g = \nabla_{\omega} R(F, D, \omega) = \sum_{f_j \in F} \sum_{(x_i, y_i) \in B \subseteq D} \nabla_{\omega} L(f_j, x_i, y_i, \epsilon_{i,j}) \tag{3}$$

In order to avoid very high variance and issues with vanishing or exploding gradient, which can occur using Eq. (3), the greedy alternative is used instead:

$$g = \frac{1}{T} \sum_{f_j \in F} \sum_{(x_i, y_i)} \sum_{t}^{T-1} \nabla_{\omega} L(f_j, x_i, y_i, \Pi_s(\epsilon^t + \delta^{t+1})) \tag{4}$$

Details on solving this optimization can be found in the original paper [24]. This learning approach is then applied to Square Attack (SA) [2] 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) networks with parameters ω_s and ω_c. Algorithm 1 describes generating adversarial images with Square Attacks where parameters are meta-learned.

Algorithm 1 Generate Adversarial Images

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

1: **for** number of epochs **do**
2: $\quad \pi^s_{\omega_s} \leftarrow trainMLP(D, SA(budget))$ (update size controller)
3: $\quad \pi^c_{\omega_c} \leftarrow trainMLP(D, SA(budget))$ (update color controller)
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**

Sobol Attribution Method aims to describe 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 Sobol attribution-based method exploits the random perturbations approach to determine complex interactions among the features and how they contribute to the outcome of $f(x)$. These random 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 vector of features x. With the set of perturbations it is possible to decompose the variance model as $\mathrm{Var}(f(x)) = \mathrm{Var}(\sum_{u \subseteq \mathcal{U}} f_u(X_u))$ where $\mathcal{U} = (1, ..., n)$, u is a subset of $\mathcal{U}$ and f_u are partial contributions of variables $X_u = (X_i)_{i \in u}$. Sobol indices are defined with the sensitivity index :

$$\mathcal{S}_u = \frac{\mathrm{Var}(f_u(X)_u)}{\mathrm{Var}(f(X))} \tag{5}$$

In this way they quantify the importance of any subset of features for the decision of the system. For their values holds $\sum_{u \in \mathcal{U}} \mathcal{S}_u = 1$. The total Sobol indices can be defined as:

$$\mathcal{S}_{T_i} = \sum_{u \subset \mathcal{U}, i \in u} \mathcal{S}_u \tag{6}$$

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. These values define each feature's intrinsic and relational impact on the model output. A low total Sobol index implies low importance for explaining the model decision. A feature weakly interacts with other features when values of the Sobol index and total are similar, while a high difference represents a strong interaction. This property of Sobol indices enables us to make the hypothesis of adversarial input detection. For estimation of Sobol indices, we use the Jensen estimator [10], which is considered an advanced one from the computation perspective. This estimator is combined with Quasi-Monte Carlo (QMC) strategy [7]. The following algorithm describes the procedure for calculating the Total Sobol index.

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: $S_{T_i} = \frac{\frac{1}{2N}\sum_{j=0}^{N}(f(A_j) - f(C_j^{(i)}))^2}{\hat{V}}$

Attribution Based Confidence (ABC) Metric is computed by importance sampling in the neighbourhood of a high-dimensional input using relative feature attributions. ABC metric constructs a generator that can sample the neighborhood of an input and observe the conformance of the model. The method does not require access to training data or any additional calibration. The concentration of features characterizes DL models. This implies that few features have high attributions for any output. 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 then in two steps: *(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. The conformance is measured by observing which values of the output did not change when the attribute changed its value.Algorithm 3 describes computing ABC metric of a DNN model on an input.

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 features $x_1, ..., x_n$ 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 samples from mutation of x_j with probability $P(x_j)$
7: Get classification output for S samples
8: $c(f, x) \leftarrow S_{conform}/S$

4 Experimental results

4.1 Implementation Details

All experiments were conducted on Google Colab Pro+ with NVIDIA T4 Tensor Core GPU and 52 GB RAM.

Dataset We have selected 1148 microscopic images from the Breast Cancer Histopathological Image Classification (BreakHis), the dataset composed of breast tumor tissue images collected from 82 patients using a 400x magnifying factor. The dataset contains two types of tumors: benign, relatively harmful, and malignant, a synonym for cancer. therefore, the dataset is associated with the image classification task into two classes. The dataset contains several types of benign tumors: tubular adenona (TA), fibroadenoma (F), adenosis (A), and phyllodes tumor (PT); and several types of malignant tumors: carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC) and papillary carcinoma (PC). We split the data from the original dataset into two sets, train and test, which correspond to 80% and 20% of the data, respectively. This split has been done in a stratified way, that is, keeping the same proportions between classes in train and test.

Classifiers We used eight pre-trained architectures for the image classification task (ResNet18, ResNet50, Inception V3, MobileNet V3, ShuffleNet, Swin Transformer, Vision Transformer, WideResnet). For breast cancer prediction, we kept the original weights of the backbone network (in some cases convolutional and in others a vision transformer) and trained a linear layer on top of them. Each model was trained on the training dataset for 500 epochs using a constant learning rate.

Adversarial attacks The Meta-RS algorithm [7], described in Section 3, was used to attack each of the eight pre-trained models. For each single classification model, controllers were meta-trained with white-box access. Advertorch package was used for adversarial training using $l_\infty - LPG$ attack with $\epsilon = 8/255$, fixed step size of 0.01, and 20 steps. MLP architectures for size and color controllers have 2 hidden layers, 10 neurons each, and ReLU activation. All correctly predicted samples were under attack during the testing. All correctly classified samples have been modified for 1000 iterations for each case.

Sobol Attribution Method [8], described in Section 3, 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 [9] parameter of the sample size was set to $S = 1000$.

[7] https://github.com/boschresearch/meta-rs
[8] https://github.com/fel-thomas/Sobol-Attribution-Method
[9] https://github.com/ma3oun/abc_metric

4.2 Classification accuracy

For trained eight classification models, we examined the classification accuracy and loss as shown in Fig. 3 during the training. The overview of the accuracy results with the test dataset for all eight models for breast cancer classification is given in Fig. 4 and Table 1. The overall highest prediction scores were achieved with Swin Transformer Network (accuracy is 0.935), and the lowest prediction score was achieved with the Inception V3 classification model (accuracy is 0.848).

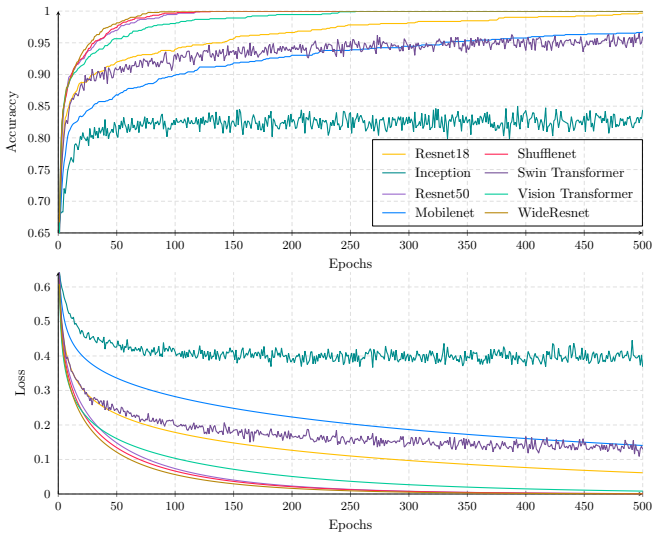

Fig. 3. Train accuracy and loss for 8 models

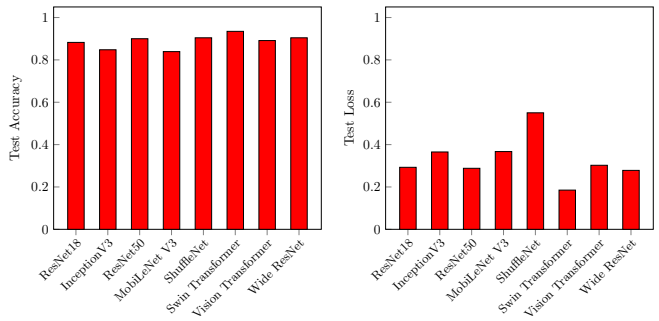

Fig. 4. Test accuracy and loss for 8 models after training for 500 epochs

4.3 Adversarial Images Generation

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. The attack accuracy is the fraction of the total number of images initially correctly classified by the model and shifted to a different class during the attack. Therefore, attack accuracy is calculated only for the previously correctly predicted labels. It is evident that adversarial attacks significantly affect prediction. The most significant decrease in accuracy is evident for the network ResNet50. An example of a successful adversarial attack is shown in Fig. 5.

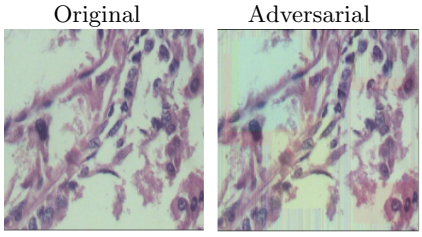

Fig. 5. An example of an image a correctly classified with the ResNet50 classifier and the corresponding successful adversarial image. High-frequency patterns and some square-shaped areas, typical patterns of the Meta-RS algorithm, can be noticed.

	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.

4.4 Detection of Adversarial Images

Explanations generated with Sobol Attribution Method represent a visual aid for a user to understand which regions of the images affected the decision-making. Examples of explanations are shown in Fig. 7 and Fig. 6 for successful attacks in the case of ResNet50 and corresponding original images. Some regions in the adversarial image are more smooth than the original, reducing the overall information and enhancing other parts' importance. It can be noticed from Sobel's attributions that prediction models do not use information from the more smooth areas and focus mainly on the parts of the images where dark spots can be found. The complexity of pathological images brings new challenges. It creates several highlighted regions within the image, rather than a few, as in the original application of the Sobol Attribution Method [5]. Hence we have developed attribution-based confidence (ABC) metric.

Table 4.4 provides ABC metric values for the original and adversarial images for all eight classification models. The computation of the ABC metric of a classification model on an input requires accurately determining conformance by sampling $S = 1000$ samples in the neighbourhood of high-dimensional inputs. The value of ABC metric in Table 4.4 is the mean value. By observing the values of ABC scores, we can draw a threshold of 0.8 for deciding whether the model was subjected to an adversarial attack or not. Fig. 8 illustrates how the ABC metric reflects the decrease in confidence under adversarial attack for all eight classification models. Fig. 9 provides examples of the final output.

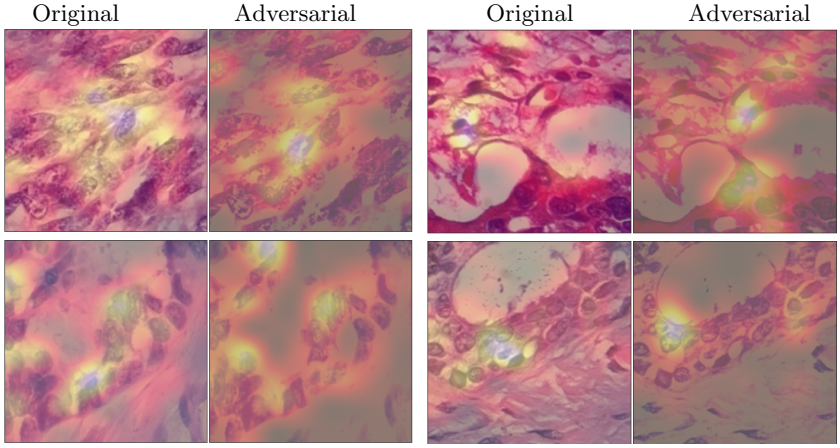

Fig. 6. Sample Sobol attribution explanations for ResNet50. Four examples of original images and their corresponding adversarial examples where attacks were successful. Sobol Attribution Method explanations are displayed on top of images highlighting the crucial regions of the image for classifier decision-making.

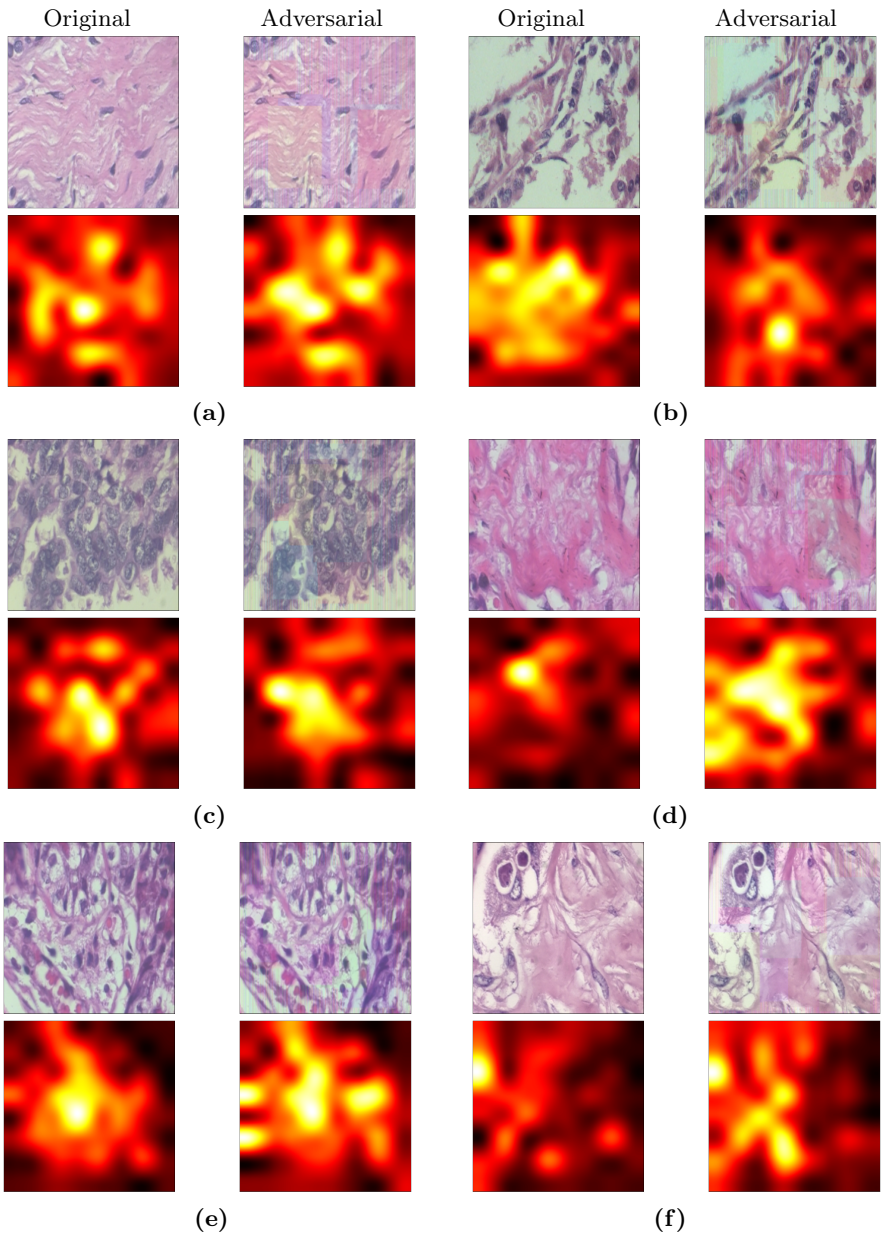

Fig. 7. Six examples of original and corresponding ResNet50 adversarial images together with sample explanations obtained with Sobol Attribution Method highlighting the importance of image regions. Several regions are being highlighted with the Sobol attribution method. This brings more challenges in deciding whether an image is original or adversarial compared to the original problem addressed by Fel et al. [5]

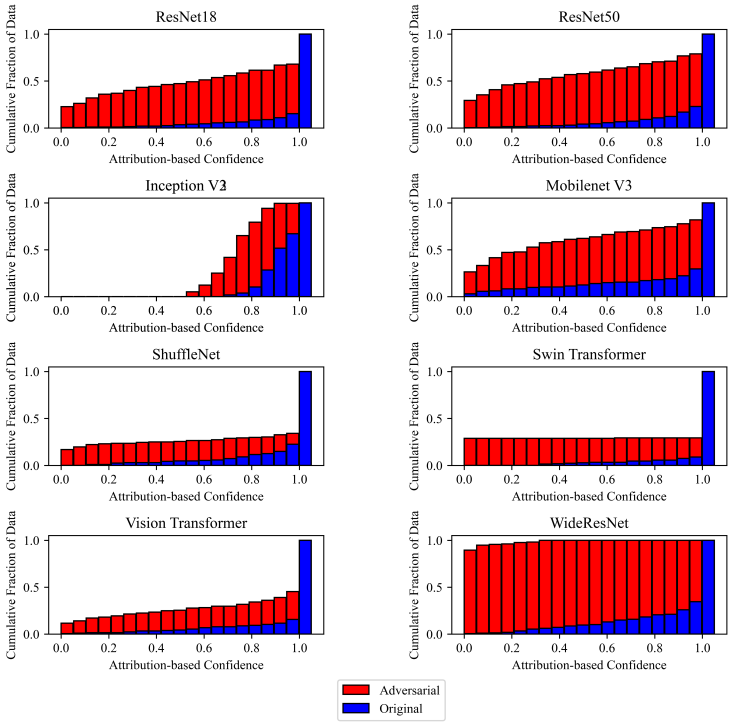

Fig. 8. Cumulative data fraction vs. ABC metric compares the original and adversarial datasets with respect to ABC metric values for different classification models. ABC metric values are high for a great fraction of data for all eight models, while for adversarial, the values are low. The most distinctive values can be recognized with the WideResNet classification model.

	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 for test and train datasets

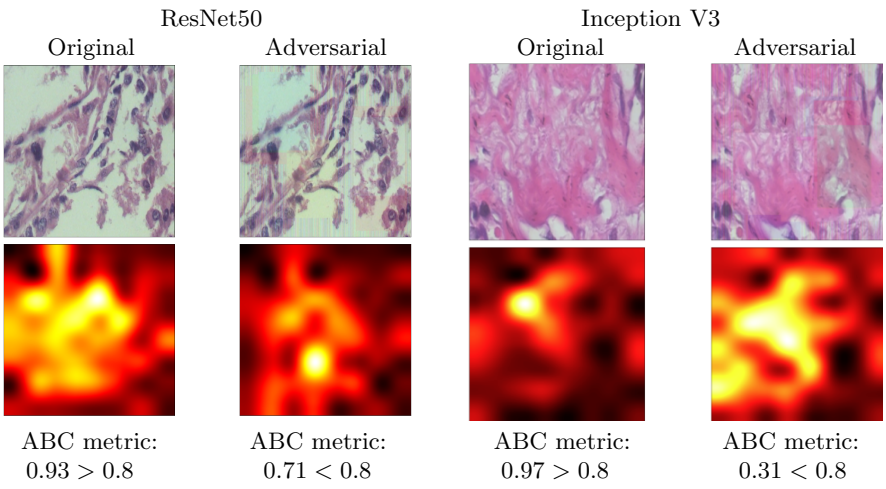

Fig. 9. ABC values are used to differentiate between original and adversarial images.

5 Conclusions

Our prediction accuracy tests on eight transfer learning-based models confirm that deep learning models have become powerful in predicting breast cancer from histopathological images. However, applying deep learning models in the medical field brings new risks and concerns, such as the possibility of adversarial attacks.

We subjected classification models to state-of-the-art robust Meta-RS attacks. The obtained adversarial images are available for public use[10]. Sobol Attribution Method [5] 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. This brings more challenges in deciding whether an image is original or adversarial compared to the original problem addressed by Fel et al. [5]. Hence we developed attribution-based confidence(ABC) metric for detecting adversarial attacks on breast histopathological images (examples in Fig. 9).

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.

In the future, the ABC metric would be used to detect adversarial attacks on histopathological oral cancer detection systems [11].

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

[11] https://www.kaggle.com/datasets/ashenafifasilkebede/dataset

References

1. Multiple Instance Captioning: Learning Representations from Histopathology Textbooks and Articles (2021)
2. Andriushchenko, M., Croce, F., Flammarion, N., Hein, M.: Square attack: A query-efficient black-box adversarial attack via random search. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIII. Lecture Notes in Computer Science, vol. 12368, pp. 484–501. Springer (2020)
3. Bortsova, G., González-Gonzalo, C., Wetstein, S.C., Dubost, F., Katramados, I., Hogeweg, L., Liefers, B., van Ginneken, B., Pluim, J.P., Veta, M., Sánchez, C.I., de Bruijne, M.: Adversarial attack vulnerability of medical image analysis systems: Unexplored factors. Medical Image Analysis **73**, 102141 (2021)
4. Das, A., Devarampati, V.K., Nair, M.S.: Nas-sgan: A semi-supervised generative adversarial network model for atypia scoring of breast cancer histopathological images. IEEE Journal of Biomedical and Health Informatics **26**(5), 2276–2287 (2022)
5. Fel, T., Cadène, R., Chalvidal, M., Cord, M., Vigouroux, D., Serre, T.: Look at the variance! efficient black-box explanations with sobol-based sensitivity analysis. In: Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. pp. 26005–26014 (2021)
6. Foote, A., Asif, A., Azam, A., Marshall-Cox, T., Rajpoot, N.M., Minhas, F.A.: Now you see it, now you dont: Adversarial vulnerabilities in computational pathology. CoRR **abs/2106.08153** (2021)
7. Gerber, M.: On integration methods based on scrambled nets of arbitrary size. Journal of Complexity **31**(6), 798–816 (2015)
8. Hashimoto, N., Fukushima, D., Koga, R., Takagi, Y., Ko, K., Kohno, K., Nakaguro, M., Nakamura, S., Hontani, H., Takeuchi, I.: Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, USA,
9. Hou, L., Agarwal, A., Samaras, D., Kurç, T.M., Gupta, R.R., Saltz, J.H.: Robust histopathology image analysis: To label or to synthesize? In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 8533–8542. Computer Vision Foundation IEEE (2019)
10. Jansen, M.J.: Analysis of variance designs for model output. Computer Physics Communications (1999)
11. Jha, S., Raj, S., Fernandes, S.L., Jha, S.K., Jha, S., Jalaian, B., Verma, G., Swami, A.: Attribution-based confidence metric for deep neural networks. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada. pp. 11826–11837 (2019)
12. Laleh, N.G., Truhn, D., Veldhuizen, G.P., Han, T., van Treeck, M., Buelow, R.D., Langer, R., Dislich, B., Boor, P., Schulz, V., Kather, J.N.: Adversarial attacks and adversarial robustness in computational pathology (2022)
13. Liu, M., Hu, L., Tang, Y., Wang, C., He, Y., Zeng, C., Lin, K., He, Z., Huo, W.: A deep learning method for breast cancer classification in the pathology images. IEEE Journal of Biomedical and Health Informatics pp. 1–8 (2022)

14. Ma, X., Niu, Y., Gu, L., Wang, Y., Zhao, Y., Bailey, J., Lu, F.: Understanding adversarial attacks on deep learning based medical image analysis systems. Pattern Recognit. **110**, 107332 (2021)
15. Marini, N., Atzori, M., Otálora, S., Marchand-Maillet, S., Müller, H.: H&e-adversarial network: a convolutional neural network to learn stain-invariant features through hematoxylin & eosin regression. In: IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, BC, Canada, October 11-17, 2021. pp. 601–610. IEEE (2021)
16. Mercan, C., Aygunes, B., Aksoy, S., Mercan, E., Shapiro, L.G., Weaver, D.L., Elmore, J.G.: Deep feature representations for variable-sized regions of interest in breast histopathology. IEEE Journal of Biomedical and Health Informatics (2021)
17. Paschali, M., Conjeti, S., Navarro, F., Navab, N.: Generalizability vs. robustness: Adversarial examples for medical imaging. CoRR **abs/1804.00504** (2018)
18. Senousy, Z., Abdelsamea, M.M., Gaber, M.M., Abdar, M., Acharya, U.R., Khosravi, A., Nahavandi, S.: Mcua: Multi-level context and uncertainty aware dynamic deep ensemble for breast cancer histology image classification. IEEE Transactions on Biomedical Engineering **69**(2), 818–829 (2022)
19. Spanhol, F.A., Oliveira, L.S., Petitjean, C., Heutte, L.: A dataset for breast cancer histopathological image classification. IEEE Transactions on Biomedical Engineering **63**(7), 1455–1462 (2016)
20. Thiagarajan, P., Khairnar, P., Ghosh, S.: Explanation and use of uncertainty quantified by bayesian neural network classifiers for breast histopathology images. IEEE Transactions on Medical Imaging **41**(4), 815–825 (2022)
21. Wei, J.W., Suriawinata, A.A., Ren, B., Liu, X., Lisovsky, M., Vaickus, L.J., Brown, C., Baker, M., Nasir-Moin, M., Tomita, N., Torresani, L., Wei, J., Hassanpour, S.: Learn like a pathologist: Curriculum learning by annotator agreement for histopathology image classification. In: IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021
22. Xu, G., Song, Z., Sun, Z., Ku, C., Yang, Z., Liu, C., Wang, S., Ma, J., Xu, W.: Camel: A weakly supervised learning framework for histopathology image segmentation (2019)
23. Yang, H., Kim, J.Y., Kim, H., Adhikari, S.P.: Guided soft attention network for classification of breast cancer histopathology images. IEEE Transactions on Medical Imaging **39**(5), 1306–1315 (2020)
24. Yatsura, M., Metzen, J., Hein, M.: Meta-learning the search distribution of black-box random search based adversarial attacks. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 30181–30195. Curran Associates, Inc. (2021)
25. Zhang, J., Ma, K., Arnam, J.S.V., Gupta, R., Saltz, J.H., Vakalopoulou, M., Samaras, D.: A joint spatial and magnification based attention framework for large scale histopathology classification. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021. pp. 3776–3784. Computer Vision Foundation / IEEE (2021)
26. Zhao, Y., Yang, F., Fang, Y., Liu, H., Zhou, N., Zhang, J., Sun, J., Yang, S., Menze, B.H., Fan, X., Yao, J.: Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020. pp. 4836–4845 (2020)