

Grad-CAM Sensitivity to Orientation Transformations in Breast Cancer Histopathological Images

Sensibilidade do Grad-CAM a Transformações na Orientação de Imagens Histopatológicas de Câncer de Mama

Felipe Moura da Silva¹, Roberta Barbosa Oliveira^{1*}

Abstract: Breast cancer is the second most prevalent among women, and early diagnosis is essential to reduce mortality. Machine learning models are being used to improve diagnostic accuracy, especially through biopsies. However, the growing complexity of neural networks has raised concerns about result interpretability, driving the adoption of explainability methods such as the Grad-CAM, which generates saliency maps in CNN models. This paper investigates the sensitivity of Grad-CAM to changes in orientation and position in breast histopathological images. CNN's ResNet50, VGG16, and Xception were evaluated, trained with the BreakHis dataset, which comprises images at different magnifications. Test samples were modified by rotations and flips to assess the robustness of both the models and the explanations. Xception achieved the best results in terms of accuracy (94.47%) and F1-score (95.85%) with 100X images, although the explanatory maps varied according to the applied transformation.

Keywords: artificial intelligence — computer vision — explainable artificial intelligence — saliency maps

Resumo: O câncer de mama é o segundo mais incidente entre mulheres, e seu diagnóstico precoce é essencial para reduzir a letalidade. Modelos de aprendizado de máquina têm sido utilizados para melhorar a precisão diagnóstica, especialmente a partir de biópsias. Contudo, a crescente complexidade das redes neurais tem gerado dúvidas quanto à interpretabilidade dos resultados, impulsionando o uso de métodos de explicabilidade como o Grad-CAM, que gera mapas de saliência em modelos CNN. Este trabalho investiga a sensibilidade do Grad-CAM frente a mudanças de orientação e posição em imagens histopatológicas da mama. Foram avaliadas as CNNs ResNet50, VGG16 e Xception, treinadas com a base BreakHis, composta por imagens de diferentes ampliações. As amostras de teste foram modificadas por rotações e inversões para testar a robustez dos modelos e explicações. O Xception obteve os melhores resultados de acurácia (94,47%) e F1-score (95,85%) com imagens 100X, embora os mapas explicativos tenham variado conforme a transformação aplicada.

Palavras-Chave: inteligência artificial — visão computacional — inteligência artificial explicável — mapas de saliência

¹ Departamento de Ciências da Computação, Universidade de Brasília (UnB), Brazil

*Corresponding author: roberta.oliveira@unb.br

DOI: <http://dx.doi.org/10.22456/2175-2745.150979> • Received: 26/10/2025 • Accepted: 24/11/2025

CC BY-NC-ND 4.0 - This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

1. Introduction

Breast cancer is the most prevalent type of cancer among female individuals in Brazil, excluding skin cancer, with 73,610 new cases per year during the 2023–2025 triennium [1].

Early diagnosis is the main strategy to reduce mortality rates and can be performed through different methods [2]. Biopsy examination is one of the possible procedures and consists of extracting a tissue sample from the suspected region and analyzing it in a laboratory. This process may result in a histopathological image, that is, a histological image with the presence of some anomaly [3]. The pathologist, through

image analysis, determines the diagnosis. However, in addition to being time-consuming, the diagnosis may be affected by various factors such as fatigue, tiredness, lack of attention, and the pathologist's experience level [4].

The increasing complexity of deep learning models, specifically Artificial Neural Networks (ANNs), has posed a challenge for researchers in the field who aim to develop techniques capable of providing a comprehensive understanding of the prediction process [5][6]. Therefore, it is preferable to develop techniques that generate explanations for a specific prediction rather than for the entire decision process of the model [7]. This also applies to Convolutional Neural

Networks (CNNs), deep learning models that use specialized layers to extract features from images [8]. CNNs have shown strong performance in image classification tasks and are thus employed in the implementation of systems designed to assist medical diagnoses [9].

The effective application of deep learning models in the medical field has faced obstacles due to the fact that doctors do not trust a decision made by a model without understanding how it reached that decision [10]. For this reason, works in the development of eXplainable Artificial Intelligence (XAI), techniques that provide an explanation of the “why” or “how” the model reached a certain decision, have gained prominence in recent years [6].

One of the most popular types of XAI is saliency map techniques, which consist of visual explanations that highlight the regions that most impacted the result of a specific prediction [11]. Because it is a visual explainability, it is often preferred by healthcare specialists [10].

The saliency map techniques can be influenced by factors such as image distortions [12]. Therefore, it is believed that these techniques may also be sensitive to certain changes in the positioning and orientation of relevant elements for image class prediction. This is an important consideration when the orientation, position, and magnification of the image being fed into the predictive model are uncertain.

The objective of this paper is to investigate one of the techniques for generating saliency map-based explanations: Grad-CAM [13]. The investigation primarily focuses on the sensitivity and invariance of this technique, given that the explanation must be effective so that pathologists and doctors can validate the predictions and obtain more accurate conclusions about disease diagnosis.

This paper is organized as follows: related works are described in Section 2; the adopted methodology is described in Section 3; the results are presented and discussed in Sections 4 and 5, respectively; finally, the conclusion and future works are addressed in Section 6.

2. Related works

The number of studies involving the application of computer vision techniques in the medical field has grown considerably. Different areas of medicine have been addressed in such studies [9]. Efforts have been made to develop models that, in addition to performing diagnosis, also provide an explanation of the result, with the intention of increasing trust in AI models, both by physicians and patients [14]. Some studies use models from the literature combined in ensembles for the task of breast cancer classification, while others have proposed models specifically designed for this task. Other studies focus on XAI techniques, proposing variations of techniques found in the literature. Finally, some aim to evaluate existing XAI techniques.

2.1 Breast cancer classification

Using the XGBoost algorithm as a classifier, fed by different convolutional networks, Maleki, Raahemi, and Nasiri [15] aimed to propose a methodology to improve the performance of deep learning models in the task of classifying breast histopathological images. The models used for feature extraction were variations of VGG and DenseNet, in addition to ResNet50. The authors concluded that the best convolutional network to use as a feature extractor is DenseNet201, where this combination achieved the best results together with XGBoost, obtaining 93.6%, 91.3%, 93.8%, and 89.1% accuracy on image sets with magnification rates of 40X, 100X, 200X, and 400X, respectively. The authors focused on binary classification of tumors, classifying them as benign or malignant.

In turn, Koshy and Anbarasi [16] built an innovative hybrid convolutional network called Levenberg–Marquardt Histology Network (LMHistNet), with the distinguishing feature of being capable of performing both binary and multiclass classification. Thus, using images from the BreakHis dataset, LMHistNet achieved very satisfactory performance, especially in binary classification, obtaining accuracies of 86% (40X), 99% (100X), 85% (200X), and 96% (400X), while in multiclass classification it obtained 71% (40X), 88% (100X), 85% (200X), and 70% (400X).

2.2 XAI techniques

Liu et al. [17] proposed both a variant of the Grad-CAM technique, called Dynamic Threshold Gradient Class Activation Mapping (DT Grad-CAM), and a new CNN model, a variation of ResNet50, called Dual-Activated Lightweight Attention ResNet50 (DALAResNet50). The DT Grad-CAM technique involves adding more steps to the Grad-CAM algorithm, namely noise generation, weighted averaging, adaptive thresholding, Otsu’s method, and morphological operations. These additional steps aim to make the technique more robust to noise and the result more interpretable and relevant for diagnosis, by improving the visualization of specific class regions. DALAResNet50 features, compared to the base model, a customized “dual attention” module, and this model achieved 98.5%, 98.7%, 97.9%, and 94.3% accuracy on images with 40X, 100X, 200X, and 400X magnification, respectively, from the BreakHis dataset, while in tests with the BATCH dataset it achieved 89.1% accuracy, and with images from Mini-DDSM, 80.07% accuracy.

Amorim et al. [18] aimed to test the reliability of the most popular saliency map techniques by evaluating the sensitivity of such techniques to image perturbations using a previously unseen strategy. The experiments focused on applying natural perturbations to the images, introducing tumors of the same or different types to assess whether and how the saliency maps would be affected. Three different scenarios were designed for testing: normal tissue perturbation applied to a normal tissue image; normal tissue perturbation applied to a tumor tissue image; and finally, tumor tissue perturbation applied

to a normal tissue image. According to the results presented, the different techniques used in the experiments proved to be sensitive to the perturbations, suggesting that saliency map techniques are a reliable form of explainability for CNNs.

2.3 Analysis of Related Works

It becomes evident that CNNs are among the most promising methods for classifying medical imaging exams due to their ability to extract features and recognize specific patterns in different regions of an image. However, it remains essential to provide robust and reliable explanations for the diagnoses produced by such models [10]. It can also be observed, from the works cited in Subsection 2.2, that explainability techniques based on saliency maps are among the most informative approaches for helping physicians and pathologists understand predicted diagnoses. Another important conclusion is that saliency map techniques are highly valuable not only for explaining diagnoses but also for enabling developers, during the training of a CNN model, to observe how feature extraction occurs, as demonstrated in the study by Maleki et al. [15].

The related works address classification and explainability tasks using saliency map methods and their analysis. However, these studies do not examine the sensitivity of such methods to spatial alterations, such as image rotation and flipping. In contrast to the approach proposed by Amorim et al. [18], the analysis of Grad-CAM sensitivity in this study did not involve perturbations to the images used in the experiments, as the goal was to analyze the variations in the maps generated for the same image under different spatial orientations. Instead, only spatial transformations, such as rotation and flipping, were applied to investigate the impact of these operations on the method's response, given that histopathological images do not have a standard orientation, unlike everyday images.

3. Methods

The trained CNN models were compared with the objective of classifying histopathological images of breast tissue obtained from the BreakHis dataset [19]. From these models, different saliency maps were generated for the same image subjected to spatial transformations, such as rotations and horizontal or vertical flips, using the Grad-CAM algorithm [13]. Subsequently, the generated maps were compared with each other, and any changes in the prediction of the tumor class were analyzed. Figure 1 presents the schematic flowchart of the proposed methodology.

The networks ResNet50 [20], VGG16 [21], and Xception [22] were adopted, as these models have a considerable capacity for feature extraction to perform image classification, being capable of recognizing patterns that help distinguish image elements [15] [23]. The XAI technique used was Grad-CAM, a post-hoc explainability technique, meaning that explanations are generated only after the model has been trained. The local nature of this technique allows for individual interpretations of each input data. Furthermore, as it is classified as an agnostic

technique, it can be applied to any CNN model, without being restricted to a specific architecture.

3.1 Dataset

The data used in this paper were obtained from the BreakHis dataset, publicly available at the UFPR website¹. BreakHis [19] contains a set of 7,909 microscopic images of breast tissue from 82 patients, captured from photographs of different regions of the slides observed under microscopes, at four different magnifications: 40X, 100X, 200X, and 400X. Figure 2 presents examples of a benign tumor and a malignant tumor at 40X magnification. In total, there are 2,480 samples of benign tumors and 5,429 malignant tumors, as shown in Table 1. All images have dimensions of 700 × 460, are in PNG format, and follow the RGB color channel standard.

Table 1. BreakHis Dataset Structure

Magnification	Benign	Malignant	Total
40X	652	1370	1995
100X	644	1437	2081
200X	623	1390	2013
400X	588	1232	1820
Total Images	2480	5429	7909

3.2 Balancing and preprocessing of datasets

To ensure better suitability of the datasets, both in the distribution of samples and in the properties of the images, it was necessary to perform procedures for dataset balancing and image preprocessing.

3.2.1 Dataset balancing

As shown in Table 1, the BreakHis dataset presents a significant imbalance, with more than twice as many malignant tumor samples compared to benign ones. This imbalance directly affects the three datasets (training, validation, and test), potentially compromising the generalization ability of the classification models, since there is a risk that the model may learn to favor the prediction of the majority class, in this case, malignant. Therefore, balancing techniques were applied exclusively to the training set, as it is the only subset used for updating the model parameters.

A method was implemented to balance the dataset using the data augmentation technique. The function developed to apply the balancing technique with data augmentation receives as arguments the dataset to be balanced, in this case, the training set, and the maximum number of samples per class. To generate new instances of the minority class, it uses a sequential layer composed of two spatial transformation operations: the first performs random flips on the horizontal and/or vertical axis, while the second applies random rotations of up to 20% relative to the angular interval 2π , that is, approximately $\pm 72^\circ$. The rotation value adopted was defined

¹<https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/>

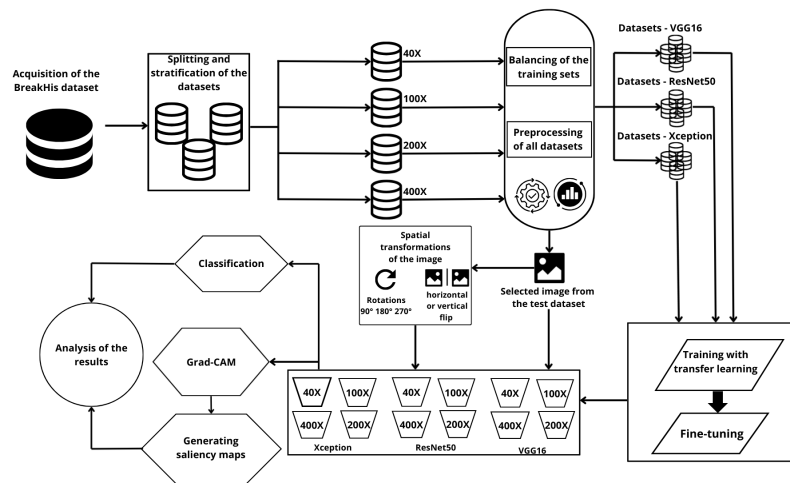


Figure 1. Methodology fluxogram

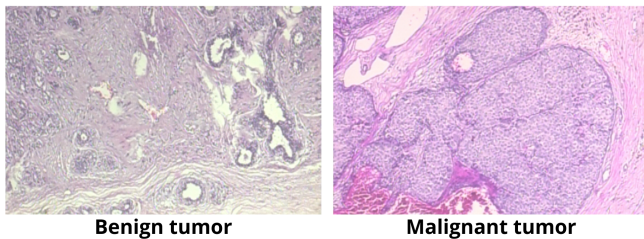


Figure 2. Examples of benign and malignant tumors at 40X magnification [24].

empirically. For the minority classes, these transformations are applied iteratively on batches of size 32 until the maximum number of samples specified in the function call is reached. The majority class does not undergo duplication.

3.2.2 Preprocessing

Due to the need for compatibility with the neural networks used in this paper, which require images with specific dimensions and pixel characteristics, it was necessary to perform sample standardization procedures. Thus, all images were resized to 224×224 using bilinear interpolation [25]. These dimensions are expected by default by the ResNet50 and VGG16 models and are compatible with Xception. Considering that each of these models requires distinct preprocessing, it was decided to generate separate versions of the datasets for each model to be trained, as well as to employ the specific methods recommended for each network, in order to facilitate the execution of the experiments and ensure the integrity of the results. In the ResNet50 and VGG16 models, the images were converted from the RGB color space to BGR, while in the Xception model, pixel values were normalized to the range between -1 and 1. These preprocessing techniques are based on the standard procedures of ImageNet [26].

3.3 Training

For model training, the transfer learning technique was adopted, using the pre-trained weights of the convolutional layers of

the base models on the ImageNet dataset, which were kept frozen during this initial phase. The models were trained for ten epochs with a high learning rate, set at 0.01. The weights corresponding to the epoch with the lowest loss function on the validation set were saved and considered the best parameters for this training stage. Table 2 presents the specific hyperparameters adopted for each architecture in this training stage. The batch size, dropout rate, activation and loss functions, and optimizer were the same for all trained models. The batch size was set to 32, a dropout rate of 0.2 was used, the adopted loss function was cross-entropy, and the optimizer defined was the Adam algorithm. The activation functions used were ReLU and sigmoid for the intermediate and output dense layers, respectively. Four versions of the ResNet50, VGG16, and Xception models were trained, one version for each magnification level: 40X, 100X, 200X, and 400X, totaling 12 models.

Table 2. Hyperparameters defined for each model during the first training stage

Model	Number of Neurons	Number of Layers	Number of Fine-tuned Layers
ResNet50	192	175	79
VGG16	480	19	4
Xception	416	132	102

Fine-tuning

In the fine-tuning stage, a specific number of convolutional layers from each model were unfrozen, as indicated in Table 2, allowing these layers to be refined for the task of tumor classification in histopathological images. In this stage, the learning rate was reduced to 0.0001, determined with the aid of tests using the Bayesian optimizer [27], and the validation loss was adopted as the objective metric. In addition to the learning rate, the Bayesian optimizer was used to define the

specific number of units in the fully connected layers of each model and the dropout rate. Training was then extended to a total of 100 epochs, using the EarlyStopping mechanism with a patience of 10 epochs to interrupt training in case of no continuous improvement in the loss metric on the validation set [28].

3.4 Saliency map using Grad-CAM

In order to produce the explanations required for the objectives of this paper, samples from the test set were randomly selected. Then, variants of these images were generated by applying different rotations or flips. For each variant, Grad-CAM was used to obtain a saliency map, which was subsequently reverted to the original orientation of the image to facilitate the comparative analysis of regions of interest under different spatial transformations.

4. Results

For the experiments, the Python 3.11 programming language was used in conjunction with specialized libraries. For the implementation and training of deep learning models, the libraries TensorFlow 2.19, Keras 3.10, and Keras Tuner 1.4 were employed. The data and image manipulation steps were conducted with the help of the libraries NumPy 2.2, Pillow 11.3.0, and OpenCV Python 4.12, while the visualization of results was enabled through Matplotlib 3.10. All experiments were carried out in a cloud-based computational environment, specifically using the Google Colab Pro service with support for the A100 GPU, which provided high performance for the tasks involved. The source code is available in a GitHub repository².

The metric used to evaluate the models in terms of classification is the F1-score, a harmonic mean of precision and recall, as shown in Equation 1. This metric provides insight into false positives and false negatives.

$$F1 - score = 2 \times \frac{(precision \times recall)}{(precision + recall)} \quad (1)$$

To evaluate the variation in the models' saliency maps, the Pearson Correlation Coefficient (CC) and the Kullback-Leibler Divergence (KL) were adopted. CC was used to measure the linear relationship between two saliency maps, interpreting them as random variables. The calculation is presented in Equation 2:

$$CC(P, Q^D) = \frac{\sigma(P, Q^D)}{\sigma(P) \times \sigma(Q^D)}, \quad (2)$$

where P is a saliency map generated by the explainability technique, and Q^D is the fixation map, while σ denotes covariance, with $\sigma(P, Q^D)$ being the covariance between P and Q [29].

KL measures how much two probability distributions diverge from each other. As it is an asymmetric measure, computing the divergence between map A and map B differs from computing the divergence between map B and map A [30]. In Equation 3, Q is the fixation map, and P is the saliency map generated by the explanation technique. Since it was not possible to obtain ground-truth maps, the maps generated from untransformed images were adopted in this paper to evaluate the variation in saliency maps as the fixation map.

$$KL(P, Q^D) = \sum_i Q_i^D \cdot \log \left(\varepsilon + \frac{Q_i^D}{\varepsilon + P_i} \right) \quad (3)$$

4.1 Binary classification

For the evaluation of the CNN models, different test sets were generated, all derived from the original set. In addition to the set containing images in their original orientation (without additional transformations), five variations were created for each of the four test subsets corresponding to the different magnification levels (40X, 100X, 200X, 400X). Each of these variations was constructed by exclusively applying one of the following specified transformations: 90° rotation, 180° rotation, 270° rotation, horizontal flip (Flip-H), and vertical flip (Flip-V). Table 3 presents the best results for each variant model, with the top-performing results among them highlighted in bold.

Table 3. Best F1-scores (%) of each variation of the trained CNN models.

Magnification	ResNet50	VGG16	Xception
40X	91.15	92.67	94.36
100X	93.52	94.70	95.85
200X	83.12	92.90	93.24
400X	90.59	90.90	88.71

4.2 Explanations using Grad-CAM

To compare the sensitivity of Grad-CAM used in the explainability of each model, CC and KL were calculated between the heatmaps of the original images and those of the altered images, as presented in Table 4. To summarize this information, the means of the results for each magnification level were calculated, along with the standard deviations to provide an idea of how much the maps deviate from the mean in each case. This allows for an understanding of how the maps at each magnification level correlate and diverge from each other. The model with the best performance evaluated in the first part of the experiments was selected to make observations regarding the sensitivity of the explanation at different magnification levels.

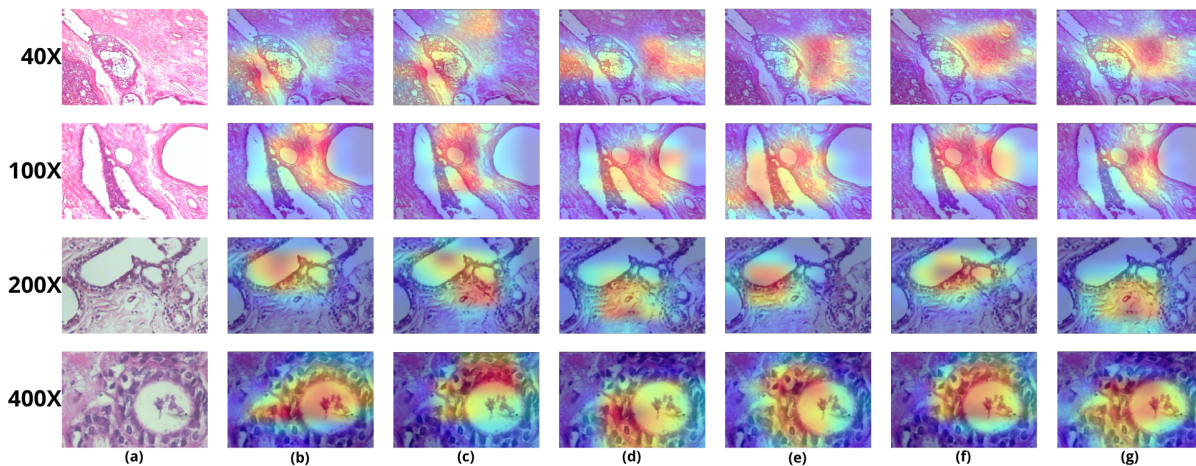
5. Discussion

The ResNet50 models performed well in tests with original images, especially at 100X magnification, but the model trained with 40X images showed a significant drop in F1-score, falling

²<https://github.com/Felipe999/Grad-CAM-Sensitivity-to-Orientation-Transformations-in-Breast-Cancer-Histopathological-Images>

Table 4. Mean and standard deviations of the Pearson Correlation Coefficient (CC) and the Kullback-Leibler Divergence (KL) of the saliency maps generated by the models across all sets with different augmentations.

	40X		100X		200X		400X	
Modelo	CC	KL	CC	KL	CC	KL	CC	KL
ResNet50	0.50	0.08	0.42	0.10	0.47	0.10	0.49	0.10
	± 0.12	± 0.02	± 0.14	± 0.02	± 0.13	± 0.02	± 0.14	± 0.03
VGG16	0.53	0.05	0.52	0.04	0.60	0.04	0.62	0.04
	± 0.10	± 0.01	± 0.12	± 0.01	± 0.14	± 0.01	± 0.14	± 0.01
Xception	0.59	0.48	0.63	0.35	0.61	0.47	0.60	0.46
	± 0.14	± 0.20	± 0.13	± 0.14	± 0.16	± 0.22	± 0.17	± 0.23

**Figure 3.** Examples of saliency maps overlaid on histopathological images of a benign tumor at 40X, 100X, 200X, and 400X magnifications, obtained from each model. (a) Original tumor images without overlaid saliency maps; (b) Saliency maps derived from the original (untransformed) images; (c–e) Saliency maps produced from the same images rotated by 90°, 180°, and 270°, respectively; (f–g) Saliency maps generated from the horizontally and vertically flipped versions of the images, respectively.

below 50%. The VGG16 models showed consistent performance, with emphasis on the model trained with 100X images, which exceeded 80% in accuracy and F1-score, demonstrating robustness against transformations. The Xception models also achieved satisfactory results, despite greater variation. The model trained with 400X images had the lowest values, but still above 80%, with the 100X model standing out.

The data in Table 4 suggest that VGG16 is the model that presents the greatest stability and the lowest variability in generating saliency maps, possibly reflecting a greater ability to identify consistent discriminative regions. Xception shows good interpretability in 100X magnification images, aligning classification and explanation quality. Finally, ResNet50 appears to be more sensitive to spatial changes, with lower consistency in visual explanations.

The Xception models, due to presenting the best classification results in most cases, were selected for the investigation involving the four image magnifications. For the observations carried out, the assumption is that if the prediction is the same for the same image under different rotations, the explanation should be similar when the heatmaps are returned to the original image orientation.

Figure 3 presents examples of saliency maps generated for

different images, each corresponding to one of the four magnifications available in the dataset, all correctly classified by the Xception model regardless of the transformations applied. Based on the previously discussed logic, it would be expected that the explanations would be similar among the variants of the same image. However, this expectation is not confirmed in certain examples. In the 40X magnification image, for instance, the saliency map in Figure 3(b) predominantly highlights the left region, while in Figures 3(e), (f), and (g), the most emphasized regions are on the right. A similar situation can be observed in the examples with 400X magnification: although there is emphasis on nearby areas among the maps, considerable variations in the location of attention are noted. Figure 3(c), for example, highlights the upper region, whereas Figure 3(d) does not show this area as relevant.

Figure 4 presents examples in which, in certain cases, the Xception model made incorrect predictions. The expectation would be that the regions of interest highlighted in the saliency maps would show significant variations in cases of classification divergence. However, when analyzing the examples of 40X magnification images, specifically Figures 4(e), where an error occurred, and Figure 4(f), with a correct prediction, it is observed that the saliency maps are quite similar, which

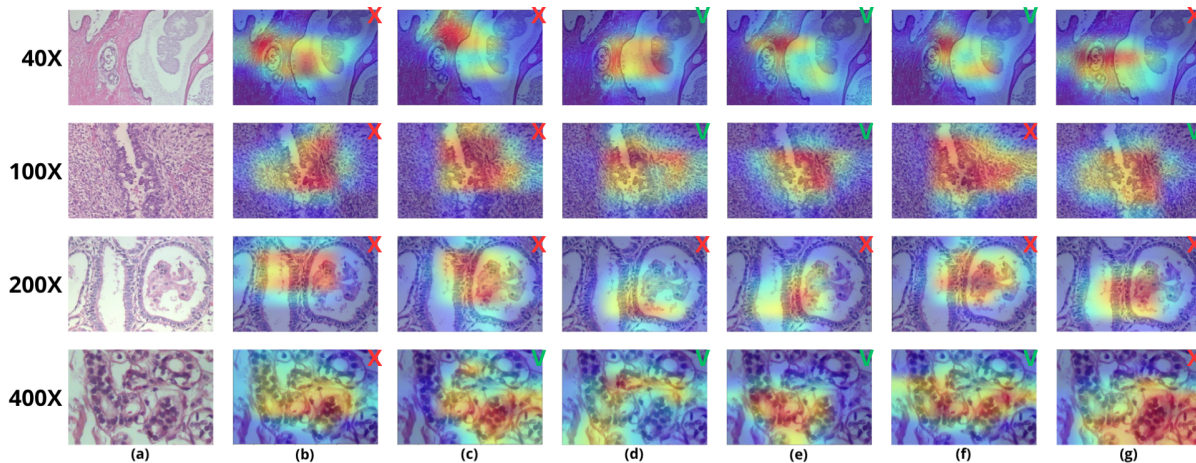


Figure 4. Examples of saliency maps overlaid on histopathological images of a malignant tumor at 40X, 100X, 200X, and 400X magnifications, generated by each model. (a) Original tumor images without overlaid saliency maps; (b) Saliency maps derived from the original (untransformed) images; (c–e) Saliency maps produced from the same images rotated by 90°, 180°, and 270°, respectively; (f–g) Saliency maps generated from horizontally and vertically flipped versions of the images, respectively.

contradicts this hypothesis. Considering that the highlighted pixels in both images received the highest weights according to the Grad-CAM algorithm, it would be expected that the predictions would also be consistent. This behavior is even more evident in the 100X magnification images, in which the central region was consistently highlighted in all variants. However, only in the example of Figure 4(g) did the model make the correct prediction. Additionally, in the image with 400X magnification, it is observed that the explanation corresponding to Figure 4(c) strongly highlights the same region emphasized in Figure 4(g), although the model produced different predictions.

The variation in the highlighted regions of the same image may result in a loss of reliability in clinical applications, even in cases where all classifications were correct. Both scenarios presented in Figures 3 and 4 may lead pathologists to question whether the models are truly capable of recognizing the relevant features necessary for an accurate diagnosis. In this work, the comparison between the generated saliency maps and the annotated regions relevant to the analysis of the samples was not possible, since the BreakHis dataset does not provide such annotations. From these annotations, it would be possible to construct ground-truth masks, enabling more rigorous quantitative evaluations and more robust conclusions regarding the effectiveness of the Grad-CAM technique. It would be interesting for future studies to consider a dataset that includes such annotations in order to obtain more conclusive results.

6. Conclusion and Future works

This paper aimed to analyze the sensitivity of Grad-CAM to changes in rotation or inversion of histopathological images, given that it is desirable for an XAI method to provide similar explanations for the same image, even when presented in different positions and qualities [12]. Xception was the best-performing model, achieving an F1-score of 95.85% on

images with 100X magnification. Assuming that correct predictions should yield correct explanations, it was observed that, in some cases, different regions of the same image were highlighted in distinct instances when that image was submitted to the model under varied orientations, both for the same model and across different models.

The models showed signs of overfitting, possibly due to the sample augmentation applied exclusively to the minority class. This balancing strategy may have led the models to memorize specific patterns of that class, reducing their generalization capacity when faced with unseen data and compromising model performance. It is possible that this issue could be mitigated through a modification of the employed algorithm, so that it also generates a controlled number of new samples from the majority class, proportional to the number of samples in the minority classes, while ensuring that the resulting dataset remains balanced. Alternatively, the dataset used, BreakHis [19], could be combined with another that contains the same tissue types and magnification levels. Another alternative would be to apply class weighting techniques to the loss functions of the trained models. This approach could eliminate the need to generate new samples, thereby reducing computational resource consumption and simplifying the model training process. Finally, cross-dataset validation could be performed to improve the generalization of the models.

For future works, a comparison with other saliency map explainability methods could be considered, such as Grad-CAM++ [31], Guided Grad-CAM [13], Occlusion [32], Integrated Gradients (IG) [33], and Guided Backprop [34]. Such a comparison would aim to evaluate which methods are more or less sensitive to spatial transformations. It is also possible to develop a new variant of Grad-CAM that is less sensitive to image variations and provides a visually enhanced explanation, where the gradients from the last convolutional layer obtained from the image under different rotations are used to

generate a more robust map. Additionally, studies involving explainability through text-generating networks [35], another approach within the field of artificial intelligence explainability, could also be considered.

Author Contributions

Felipe Moura da Silva: Conducted the literature review, implemented the models, performed the experiments, analyzed the results, and wrote the manuscript and the work that inspired this paper [36].

Roberta Barbosa Oliveira: Supervised the research, provided guidance on methodology and experimental design, contributed to the analysis and interpretation of results, and revised the manuscript.

References

- [1] MARMITT, N. et al. Cenário das publicações sobre perfil epidemiológico e clínico do câncer de mama: Revisão de escopo. *LUMEN ET VIRTUS*, v. 16, n. 48, p. 6027–6048, maio 2025. ISSN 2177-2789. Disponível em: <https://periodicos.newsciencepubl.com/LEV/article/view/5511>.
- [2] RAMALHO, R. B. et al. Diagnóstico e tratamento do câncer de mama: uma revisão de literatura. *Brazilian Journal of Implantology and Health Sciences*, v. 6, n. 8, p. 1040–1050, ago. 2024. Disponível em: <https://bjih.emnuvens.com.br/bjih/article/view/2866>.
- [3] HOUSSEIN, E. H. et al. Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review. *Expert Systems with Applications*, v. 167, p. 114161, abr. 2021. ISSN 0957-4174. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0957417420309015>.
- [4] YAN, R. et al. Breast cancer histopathological image classification using a hybrid deep neural network. *Methods*, v. 173, p. 52–60, fev. 2020. ISSN 10462023. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S1046202319300349>.
- [5] ČERNEVIČIENĖ, J.; KABAŠINSKAS, A. Explainable artificial intelligence (XAI) in finance: a systematic literature review. *Artificial Intelligence Review*, v. 57, n. 8, p. 216, jul. 2024. ISSN 1573-7462. Disponível em: <https://doi.org/10.1007/s10462-024-10854-8>.
- [6] YANG, G.; YE, Q.; XIA, J. Unbox the black-box for the medical explainable ai via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Information Fusion*, v. 77, p. 29–52, 2022. ISSN 1566-2535. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1566253521001597>.
- [7] ARRIETA, A. B. et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *arXiv preprint arXiv:1910.10045*, 2019. Disponível em: <https://doi.org/10.48550/arXiv.1910.10045>.
- [8] ZHAO, X. et al. A review of convolutional neural networks in computer vision. *Artificial Intelligence Review*, v. 57, n. 4, p. 99, mar. 2024. ISSN 1573-7462. Disponível em: <https://doi.org/10.1007/s10462-024-10721-6>.
- [9] ESTEVA, A. et al. Deep learning-enabled medical computer vision. *npj Digital Medicine*, v. 4, n. 1, p. 1–9, jan. 2021. ISSN 2398-6352. Disponível em: <https://www.nature.com/articles/s41746-020-00376-2>.
- [10] PATRÍCIO, C.; NEVES, J. C.; TEIXEIRA, L. F. Explainable Deep Learning Methods in Medical Image Classification: A Survey. *ACM Computing Surveys*, v. 56, n. 4, p. 1–41, abr. 2024. ISSN 0360-0300, 1557-7341. Disponível em: <https://dl.acm.org/doi/10.1145/3625287>.
- [11] AMORIM, J. P. et al. Interpreting Deep Machine Learning Models: An Easy Guide for Oncologists. *IEEE Reviews in Biomedical Engineering*, v. 16, p. 192–207, 2023. ISSN 1941-1189. Disponível em: <https://ieeexplore.ieee.org/document/9629296>.
- [12] LIU, Y. et al. Optimized Dropkey-Based Grad-CAM: Toward Accurate Image Feature Localization. *Sensors*, v. 23, n. 20, p. 8351, jan. 2023. ISSN 1424-8220. Disponível em: <https://www.mdpi.com/1424-8220/23/20/8351>.
- [13] SELVARAJU, R. R. et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. [s.n.], 2017. p. 618–626. Disponível em: <https://ieeexplore.ieee.org/document/8237336>.
- [14] BAI, S. et al. *Breast Cancer Diagnosis: A Comprehensive Exploration of Explainable Artificial Intelligence (XAI) Techniques*. arXiv, 2024. Disponível em: <http://arxiv.org/abs/2406.00532>.
- [15] MALEKI, A.; RAAHEMI, M.; NASIRI, H. Breast cancer diagnosis from histopathology images using deep neural network and XGBoost. *Biomedical Signal Processing and Control*, v. 86, p. 105152, set. 2023. ISSN 17468094. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S1746809423005852>.
- [16] KOSHY, S. S.; ANBARASI, L. J. LMHistNet: Levenberg–Marquardt Based Deep Neural Network for Classification of Breast Cancer Histopathological Images. *IEEE Access*, v. 12, p. 52051–52066, 2024. ISSN 2169-3536. Disponível em: <https://ieeexplore.ieee.org/document/10491266/>.
- [17] LIU, S.; HIMEL, G. M. S.; WANG, J. Breast Cancer Classification With Enhanced Interpretability: DALAResNet50 and DT Grad-CAM. *IEEE Access*, v. 12, p. 196647–196659, 2024. ISSN 2169-3536. Disponível em: <https://ieeexplore.ieee.org/document/10810414/?number=10810414>.

- [18] AMORIM, J. P. et al. Evaluating the faithfulness of saliency maps in explaining deep learning models using realistic perturbations. *Information Processing & Management*, v. 60, n. 2, p. 103225, mar. 2023. ISSN 03064573. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0306457322003260>).
- [19] SPANHOL, F. A. et al. A Dataset for Breast Cancer Histopathological Image Classification. *IEEE Transactions on Biomedical Engineering*, v. 63, n. 7, p. 1455–1462, jul. 2016. ISSN 0018-9294, 1558-2531. Disponível em: <http://ieeexplore.ieee.org/document/7312934/>.
- [20] HE, K. et al. *Deep Residual Learning for Image Recognition*. arXiv, 2015. Disponível em: <http://arxiv.org/abs/1512.03385>.
- [21] SIMONYAN, K.; ZISSERMAN, A. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. arXiv, 2015. Disponível em: <http://arxiv.org/abs/1409.1556>.
- [22] CHOLLET, F. *Xception: Deep Learning with Depthwise Separable Convolutions*. arXiv, 2017. Disponível em: <http://arxiv.org/abs/1610.02357>.
- [23] AL., A. et. Breast cancer detection and classification using deep learning xception algorithm. *International Journal of Advanced Computer Science and Applications*, The Science and Information Organization, v. 13, n. 7, 2022. Disponível em: <https://dx.doi.org/10.14569/IJACSA.2022.0130729>.
- [24] BREAST Cancer Histopathological Database (BreakHis) – Laboratório Visão Robótica e Imagem. Disponível em: <https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/>.
- [25] Cambridge in Colour. *Digital Image Interpolation*. 2020. Disponível em: <https://www.cambridgeincolour.com/tutorials/image-interpolation.htm>.
- [26] DENG, J. et al. Imagenet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2009. p. 248–255.
- [27] FRAZIER, P. I. *A Tutorial on Bayesian Optimization*. 2018. Disponível em: <https://arxiv.org/abs/1807.02811>.
- [28] PRECHELT, L. Early stopping — but when? In: ORR, G. B.; MÜLLER, K.-R. (Ed.). *Neural Networks: Tricks of the Trade*. Berlin, Heidelberg: Springer, 1998, (Lecture Notes in Computer Science, v. 1524). p. 55–69. Disponível em: https://doi.org/10.1007/3-540-49430-8_3.
- [29] BYLINSKII, Z. et al. *What do different evaluation metrics tell us about saliency models?* arXiv, 2017. Disponível em: <http://arxiv.org/abs/1604.03605>.
- [30] LOPATECKI, J. *KL Divergence: When To Use Kullback-Leibler divergence*. Disponível em: <https://arize.com/blog-course/kl-divergence/>.
- [31] CHATTOPADHYAY, A. et al. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*. [s.n.], 2018. p. 839–847. Disponível em: <https://arxiv.org/abs/1710.11063>.
- [32] ZEILER, M. D.; FERGUS, R. Visualizing and understanding convolutional networks. In: *European Conference on Computer Vision (ECCV)*. Springer, 2014. p. 818–833. Disponível em: <https://arxiv.org/abs/1311.2901>.
- [33] SUNDARARAJAN, M.; TALY, A.; YAN, Q. Axiomatic attribution for deep networks. In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*. [s.n.], 2017. p. 3319–3328. Disponível em: <https://arxiv.org/abs/1703.01365>.
- [34] SPRINGENBERG, J. T. et al. *Striving for Simplicity: The All Convolutional Net*. 2015. Disponível em: <https://arxiv.org/abs/1412.6806>.
- [35] POLI, J.-P.; OUERDANE, W.; PIERRARD, R. Generation of textual explanations in xai: the case of semantic annotation. In: *2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. [S.l.: s.n.], 2021. p. 1–6.
- [36] SILVA, F. M. *Explicabilidade Visual de Redes Convolucionais em Diagnóstico Histopatológico de Câncer de Mama: Sensibilidade do Grad-CAM a Transformações na Orientação das Imagens - Trabalho de Conclusão de Curso - Computação (licenciatura) - Departamento de Ciência da Computação - Universidade de Brasília*. 2025.