

Impact of Anatomical Positioning Markers on Breast Cancer Detection Thermography

João Augusto Pilato de Castro, Fabrício Araújo Filgueiras, Luíza Machado Costa de Nascimento, Heder Soares Bernardino(✉), and Saulo Moraes Villela

Federal University of Juiz de Fora, Juiz de Fora, MG, Brazil
pilato.joao@estudante.ufjf.br, fabricio.filgueiras@estudante.ufjf.br, luiza.machado@estudante.ufjf.br, heder.bernardino@ufjf.br, saulo.moraes@ufjf.br

Abstract. In the literature on thermographic breast image classification, methods of pre-processing are used to focus on relevant information. Segmentation of the region of interest and elimination of noise are examples widely adopted. However, such approaches can eliminate relevant contextual features used on learning process. In the datasets utilized by the majority of studies, one relevant element is a marker, often discarded by those techniques. This study investigates the influence of this marker on breast cancer identification, through combination of different aspects influenced by the marker. We studied the effects of its presence, as well as the impact of different methods of pre-processing images. Results show that the markers can improve perfomance, and its removal or displacement can impact directly on learning process. These findings suggest that the marker collaborates in the evaluation of the entire image field, focusing on discriminative features that indicates presence of cancer, demonstrating that external elements should not be treated as noise, but may contribute to interpreting spatial patterns in thermographic data.

1 Introduction

Breast cancer represents one of the major global public health concerns. According to data from the International Agency for Research on Cancer (IARC)[1], it is estimated that approximately one in every twenty women will be diagnosed with the disease over their lifetime. Projections also suggest that this scenario may worsen in the coming decades. According to Nature Medicine [1], if current trends are maintained, the annual number of new cases could reach 3.2 million by 2050, with an estimated 1.1 million deaths.

In this context, thermography has been increasingly explored as a complementary approach for breast cancer screening. As a non-invasive technique, it enables the mapping of thermal patterns associated with tumor metabolic activity without requiring physical contact [2,3]. In addition, its low operational cost,

[1] https://www.iarc.who.int.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16768, pp. 126–140, 2027.
https://doi.org/10.1007/978-3-032-30488-9_8

combined with the absence of ionizing radiation [4], makes it a suitable option for early detection, especially in settings with limited resources.

Despite these advantages, the clinical interpretation of thermographic images remains challenging [5], as the observed thermal patterns are often subtle and influenced by multiple physiological factors. For this reason, deep learning models have been used to identify patterns that are difficult to detect through visual inspection alone [6,7].

Recent advances in the field have focused mainly on architectural improvements. Even so, many computational pipelines still rely on aggressive preprocessing strategies based on region of interest (ROI) segmentation to remove background noise and emphasize the relevant anatomical region [8]. In practice, this means that external visual elements are often treated as irrelevant to learning process. As a result, acquisition related artifacts and reference markers are commonly removed, discarding information that may contribute to model performance.

This assumption becomes more problematic when considering that thermographic datasets often contain structured visual patterns introduced during image acquisition. These patterns may include methodological artifacts, such as segmentation masks, as well as physical elements, such as anatomical positioning markers, which create localized thermal discontinuities. Although these elements are often discussed in the context of shortcut learning [9], where models rely on non-causal correlations, their presence does not always imply a spurious effect. In some cases, acquisition related elements may provide spatial or geometric references that interact with the region of interest.

This issue becomes even more relevant because the literature relies on a small number of public datasets, especially the Visual Lab Database [10], which has been widely used in breast thermography research [11,12]. When many studies depend on the same dataset, recurring acquisition patterns and artifacts are more likely to influence model behavior in a systematic way. This makes controlled investigations into their role in the learning process especially important.

Preliminary experiments with the Visual Lab dataset showed that models trained on original images containing the anatomical positioning marker consistently performed better than those trained on images in which the marker was removed or indirectly excluded through segmentation. This result is counterintuitive, since conventional preprocessing strategies would normally suggest that removing external elements should improve model generalization. These observations indicate that such elements may do more than introduce noise, as they can function as contextual references that affect how models interpret thermal patterns.

Based on these observations, this study investigates the role of non-physiological visual elements in deep learning models for breast cancer classification. Specifically, we analyze the impact of the anatomical positioning marker (TAG) under different preprocessing conditions, including marker removal, spatial relocation, and normalization of its physical properties. The goal is to determine whether these elements behave as spurious shortcuts or as meaningful contextual cues that contribute to model performance.

2 Related Work

The application of neural models for the classification of breast thermograms is not a new idea, recent systematic reviews indicate a wide range of studies in this area with very promising results. Looking at the methods used and the results obtained, one thing is clear CNNs (Convolutional Neural Networks) are the most prevalent models, almost always along with Transfer Learning. Although classic architectures such as VGG [13] and ResNet [14] work well yield strong results, newer models that prioritize computational efficiency, such as EfficientNet and MobileNet, appear far less frequently [15]. This suggests that the field has become stuck in what is already known and has failed to explore what is new.

Another problem lies in the fact that the overwhelming majority of the literature is dependent on a single dataset. More than 77% of studies are based solely on the Visual Lab (DMR-IR) image set from the Federal Fluminense University. This creates what some authors call "data hegemony," which, while standardizing comparisons, imposes technical and demographic biases. Since this dataset has become the standard in the field, we must ask ourselves: what works within it, does it work outside of it? What kind of variation completely changes the performance of the models? This type of analysis is lacking.

In the context of preprocessing, the literature largely focuses on basic procedures: resizing, normalizing, and augmenting data. This helps to circumvent overfitting, mainly because the dataset is small, but it is not as interesting as a case study because these techniques ignore details that come from the image acquisition itself, such as anatomical markers, relevant background regions, and non-physiological patterns.

Bringing all this information together from the literature, it becomes clear that there is a significant gap in the literature regarding the analysis of visual biases and acquisition methods that affect the performance and interpretability of CNN-based models in breast thermography. This work seeks to question this limitation and investigate, in a controlled manner, how these factors impact the ability of models to generalize to new data, comparing their reliability and the results obtained.

3 Materials and Methods

In order to systematically investigate the influence of the marker will describe the following in four parts: the dataset used and its particularities, how the preprocessing was done, the models and architectures used, and finally the experimental protocol adopted.

3.1 Dataset

This study uses the dataset Visual Lab Database [10], a public repository widely used in the literature on breast thermography research. It consists of several thermographic images captured with a fixed protocol, at five different angles,

including frontal, left lateral 45°, left lateral 90°, right lateral 45°, and right lateral 90° views.

For our study, only frontal images were considered to ensure greater geometric consistency between samples and to avoid unnecessary variability from different angles. Therefore, the main dataset used contains only frontal images, all captured under controlled conditions.

An interesting detail in the images is that they have a square marker attached to the patient's chest, here called an anatomical positioning marker (TAG), which serves as a geometric reference during the image acquisition process, appearing in virtually all images and always in the same style (Fig. 1).

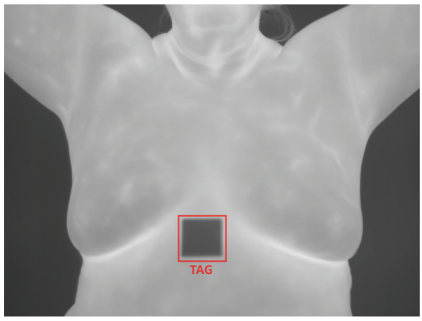


Fig. 1. Example of an image from the dataset.

3.2 Preprocessing and Data Augmentation

In the domain of machine learning applied to computer vision, preprocessing and data augmentation constitute fundamental stages for dataset preparation and enrichment [16], serving distinct methodological purposes.

The preprocessing stage encompasses a more mechanical concept, in which standardized transformations are applied to the raw image, such as resizing and normalization on the temperature scale [17], with the goal of leaving all samples in a canonical format and aligned for the model to process. In data augmentation, the idea is to synthetically generate new samples through controlled random transformations, rotating the images slightly, for example. This increases the diversity of the training set and acts as an implicit regularization [18] during network training.

3.3 Model Architectures and Training

For the main analysis architecture, VGG-16 (Visual Geometry Group) [13] was chosen, a deep convolutional network architecture created by Simonyan and Zisserman in 2014 that is characterized by systematically using small 3×3 filters

and organizing the layers in a more homogeneous way. There are two more common versions, VGG-16 and VGG-19 (16 and 19 weighted layers, respectively). What is striking about this neural network is its conceptual simplicity and the proof that a critical factor in obtaining good results is the number of layers; the more layers, the better the hierarchical representations it learns.

4 Methodology

Prior to defining the formal evaluation protocol, preliminary experiments were conducted to visually assess model behavior under different preprocessing strategies.

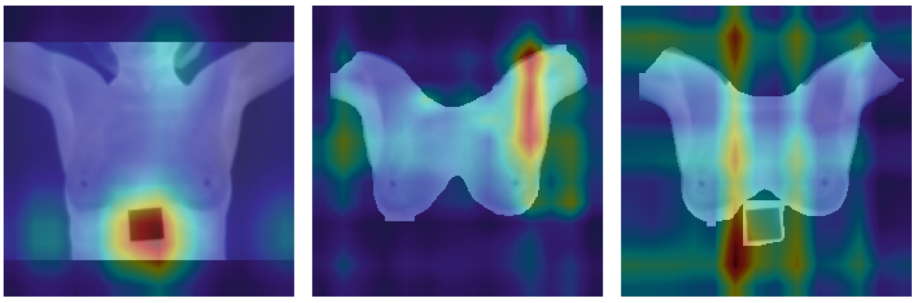
These initial observations suggested that model performance is influenced by factors beyond the region of interest, motivating a more controlled and systematic experimental design.

4.1 Experimental Protocols and Variations

The experimental design was guided by preliminary qualitative observations of model behavior under different preprocessing strategies. In particular, models trained on original images consistently outperformed those trained on segmented regions.

To better understand this discrepancy, activation maps (EigenCAM) were analyzed, revealing that the networks frequently concentrated attention on the anatomical positioning marker (TAG) in the original images.

Comparative observations across original, segmented, and segmented-with-marker images indicated that the presence of the TAG systematically alters the model's attention patterns, as illustrated by the activation maps in Fig. 2. This effect was especially evident in correctly classified healthy cases, where no strong pathological thermal patterns are present.



(a) Original image (with TAG) (b) Segmented image (without TAG) (c) Segmented image with preserved TAG

Fig. 2. EigenCAM activation maps.

These findings suggested that removing the marker through segmentation may eliminate not only irrelevant background information, but also a potentially useful spatial reference. To investigate this hypothesis, we designed a set of controlled experimental configurations aimed at isolating the individual contributions of segmentation and marker presence.

To systematically evaluate these effects, we defined a set of experimental configurations, described in the following subsection.

4.2 Creation of Manipulated Subsets and Justification

The Size, Position, and Temperature Normalized subsets were derived from the original dataset (ORIG). The TAG region was manually delineated using Adobe Photoshop, producing high-precision masks that ensured geometric consistency across all subsequent manipulations.

Breast-Segmented Set (SEG). The segmented subset was defined as a complementary negative control. Images were obtained through the automatic segmentation of the breast region using a previously trained YOLO-Seg model, subsequently specialized with breast images and masks from the same dataset, removing extra-mammary thoracic tissue and consequently the TAG. This subset allows for the assessment of the impact of the strict restriction on the region of interest (ROI) on model performance in a scenario where contextual information external to the breast is eliminated.

Breast-Segmented Set with Preserved TAG (SEG+TAG). In order to dissociate the effect of ROI segmentation from the effect of marker removal, a subset was created in which automatic breast segmentation was performed while explicitly preserving the region corresponding to the TAG. For this purpose, breast segmentation masks obtained by the YOLO-Seg model were combined with manual marker masks, resulting in images where only breast tissue and the TAG are retained.

This configuration allows for the evaluation of whether the performance degradation observed in segmented scenarios arises from marker removal or from the spatial restriction to the ROI itself. Thus, the SEG+TAG subset works as an intermediate control, allowing us to analyze the effect of segmentation without removing the marker.

TAG-Removed Set (WT Without TAG). In the WT. subset, the TAG was completely removed from the original images. The marker was removed using Photoshop's Healing Brush tool, which fills the region by sampling texture and temperature from the surrounding thoracic area. This subset represents an extreme scenario in which the marker is entirely eliminated, allowing assessment of the effect of its absence on the classification process.

TAG Size-Normalized Set (FS Fixed Size). To analyze the effect of marker size, a dataset was created in which the TAG size was fixed. First, the smallest marker dimension present in the original set was identified (28×28 pixels). For each image, a square region of this size was cropped from within the original TAG. Subsequently, the original marker was removed using the Healing Brush tool, and the cropped region was reinserted in the same original position. This manipulation eliminates variability in marker size, enabling the evaluation of whether TAG scale influences the model learning process.

TAG Moved Set (MV Moved). To isolate the effect of spatial marker location, the TAG was repositioned in all images to a fixed and standardized location outside the thoracic region, in the right corner of the image. For this purpose, the marker was manually cropped, the original area reconstructed using the Healing Brush tool, and the TAG pasted into the new position. This subset allows for the evaluation of whether the anatomical position of the marker influences learning, independently of its shape or thermal signature.

TAG Temperature-Normalized Set (FT Fixed Temperature). To analyze the influence of the marker's thermal signature, pixel values corresponding to the TAG region were homogenized. A single TAG was selected as a canonical reference, and the intensity value at its central point (27.87° C) was sampled using the Eyedropper tool. This value was then applied to all pixels within the TAG mask across the remaining images, replacing the original thermal signatures. This manipulation removes thermal variability within the marker, making it possible to evaluate whether the model is sensitive to its thermal pattern.

4.3 Preprocessing, Training, and Validation Protocol

To ensure that differences between experimental subsets are due only to TAG manipulations, a standardized protocol for preprocessing, training, and evaluation was adopted. The pipeline described below, illustrated in Fig. 3, was applied identically across all configurations, keeping the model architecture, training parameters, and validation criteria fixed. This allows direct comparison between the evaluated scenarios.

Data Partitioning and Normalization. The data were split into training (60%), validation (20%), and test (20%) sets, keeping patient-level separation between them. This avoids information leakage and ensures that the models are evaluated on unseen patients.

Image normalization was performed using min–max with a $[0, 1]$ range. To prevent data leakage, the minimum and maximum values used for normalization were computed exclusively from the training set and applied to the corresponding validation and test sets.

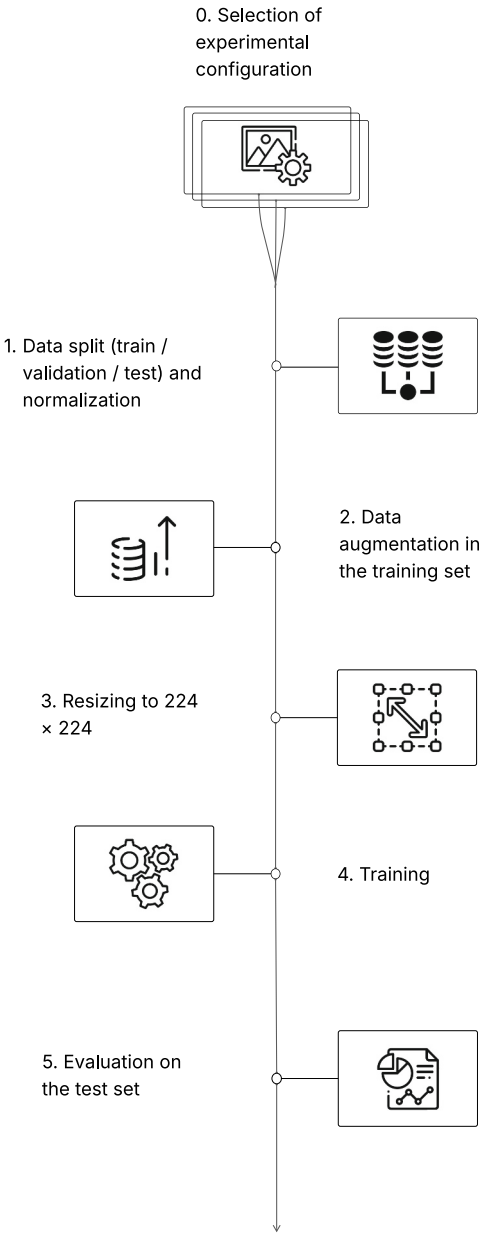


Fig. 3. Overview of the experimental pipeline.

Resizing and Data Augmentation. To standardize network input dimensions and reduce computational costs without introducing geometric distortions, all images were resized to a resolution of 224 × 224 pixels using letterbox resizing

with black padding. This approach preserves the original aspect ratio, preventing deformations.

To mitigate overfitting and improve model generalization, data augmentation techniques were applied exclusively to the training set. Transformations included small-magnitude rotations, zoom, brightness variations, and horizontal flipping, which were applied with a 50% probability in conjunction with the other transformations.

Model Architecture and Training Configuration. Experiments were conducted using a VGG-16-based Convolutional Neural Network architecture, selected due to its extensive validation in the literature for medical and thermographic image classification tasks. This choice makes it easier to compare our results with previous studies and helps focus the analysis on preprocessing and marker manipulation rather than architectural changes.

The network consists of five convolutional blocks with 3×3 filters, followed by max pooling and batch normalization. The fully connected layers include two dense layers with 4096 neurons each, interleaved with dropout at a rate of 0.5 to reduce overfitting. The output layer has a single neuron with a sigmoid activation function for binary classification (healthy vs diseased).

Training and Model Selection. Training was performed using binary cross-entropy loss and the Adam optimizer, with an initial learning rate of 10^{-4}. To improve computational efficiency, experiments were conducted under a mixed precision policy and keeping the numerical stability of the loss function.

Each model was trained for up to 500 epochs, with early stopping monitoring validation loss, using a patience of 50 epochs and a minimum improvement threshold of 0.01. The model achieving the best validation performance was stored and subsequently used for evaluation on the test set.

In scenarios involving segmentation, a pre-trained YOLOv8-Seg model was previously specialized using breast images and masks for ROI segmentation and was employed to automatically segment images in the training, validation, and test sets according to the evaluated experimental configuration.

5 Computational Experiments

This section presents the computational experiments designed to evaluate, in a controlled and systematic manner, the impact of different preprocessing strategies and Anatomical Positioning Marker (TAG) manipulations on model performance. The experiments are structured to ensure direct comparability across configurations, allowing the isolation of the effects associated with each modification. We adopted multiple repetitions in each experiment to evaluate the significance of each configuration with or without the marker, which are detailed in the following subsections.

5.1 Experimental Protocol

For the purpose of comparison, initially the models were trained using the original images, without extraction of features, in other words, full images. Subsequently, the tag configurations were modified to simulate different scenarios, including the marker removed artificially or as a consequence of ROI segmentation.

The same VGG-16-based architecture was adopted, and the data partitioning, training parameters, and preprocessing steps remained the same. The goal was to observe how TAG manipulation and dataset arrangement affected the training results.

5.2 Evaluation Strategy and Repetitions

To consistently evaluate each experiment, 30 independent runs were performed for each configuration, totaling 210 executions. Distinct random seeds were adopted for both data partitioning and network weight initialization.

The evaluation metrics at the end of each run were ROC_AUC, F1-score, Recall, Precision, and Accuracy, using a fixed decision threshold of 0.5. The final results of each experiment were obtained from the model's binary predictions, using the mean of each metric. This approach was applied to reduce the influence associated with the weight initialization and data partitioning of each independent run.

5.3 Statistical Analysis Procedure

A number of independent runs were performed during the experiments. To determine whether the observed differences among the means of the experimental configurations are statistically significant, a multi-stage statistical analysis procedure was adopted. First, the normality of the metric distributions was assessed using the Shapiro–Wilk test. Results indicated that the Accuracy metric presented at least one group violating the normality assumption, whereas ROC_AUC, F1-score, Recall, and Precision satisfied this assumption across all evaluated configurations.

So, analysis of variance was applied exclusively to these four metrics, using a one-way Welch ANOVA, which is suitable for independent samples and does not require homogeneity of variances. A significance level of 5% was used.

Following the rejection of the ANOVA null hypothesis, post hoc comparisons were performed using the Games–Howell test to identify which pairs of configurations differed statistically.

6 Results and Discussion

Here, we discuss the experimental results obtained from the controlled manipulation of TAG in thermographic breast images. The analysis is conducted on

two levels: (i) a global statistical assessment to verify whether the preprocessing differences significantly affect the model's performance, and (ii) a pairwise analysis to identify which specific TAG features are responsible for the observed differences.

6.1 Descriptive Performance Analysis

The means and standard deviations of the performance metrics across the 30 independent runs for each experimental setup are summarized in Table 1. A consistent performance hierarchy pattern is observed across all evaluated metrics.

Table 1. Means and standard deviations of performance metrics for different configurations. FT (TAG Temperature-Normalized Set with Fixed Temperature) and SEG+TAG (Breast-Segmented Set with Preserved TAG).

Configuration	Accuracy	Precision	Recall	F1-Score	ROC_AUC
FT	0.875 ± 0.039	0.839 ± 0.062	0.836 ± 0.097	0.833 ± 0.053	0.955 ± 0.023
Original	0.871 ± 0.045	0.823 ± 0.081	0.847 ± 0.073	0.832 ± 0.063	0.944 ± 0.024
TAG Size-Normalized	0.851 ± 0.043	0.797 ± 0.071	0.827 ± 0.087	0.807 ± 0.057	0.939 ± 0.024
SEG + TAG	0.853 ± 0.048	0.808 ± 0.077	0.806 ± 0.128	0.799 ± 0.075	0.924 ± 0.033
TAG-Moved	0.831 ± 0.039	0.780 ± 0.068	0.775 ± 0.093	0.773 ± 0.055	0.899 ± 0.039
Breast-Segmented	0.831 ± 0.053	0.791 ± 0.103	0.755 ± 0.107	0.766 ± 0.080	0.906 ± 0.040
TAG-Removed	0.825 ± 0.042	0.774 ± 0.082	0.764 ± 0.106	0.762 ± 0.060	0.905 ± 0.040

The original dataset (ORIG) and the dataset with the Fixed Temperature Marker (FT) showed the highest average values across all evaluated metrics. Conversely, the configurations in which the marker was removed (WT), spatially shifted (MV), or implicitly eliminated due to breast segmentation (SEG) consistently exhibited lower performance. Intermediate results were observed for configurations that preserve the marker under controlled conditions, such as fixed size (FS) or segmentation with marker preservation (SEG+TAG).

This initial observation suggests that the presence of TAG is positively associated with the performance of the models. However, descriptive statistics alone are not sufficient to determine whether these differences are statistically significant or if they are simply due to the inherent stochastic variability in training deep neural networks.

6.2 Global Statistical Analysis: Does the Marker Make a Difference?

To assess whether the differences between the results are statistically significant, a one-way ANOVA with Welch's correction was applied to the ROC_AUC, F1-score, Recall, and Precision values. This choice is justified by the independence of the samples and the absence of the homogeneity of variances assumption (Sect. 5).

In Table 2, the ANOVA results indicate statistically significant differences between the configurations for all four analyzed metrics ($p < 0.05$). ROC_AUC ($\eta^2 \approx 0.48$) and F1-score ($\eta^2 \approx 0.32$) show moderate to high effect sizes, suggesting a strong influence of the evaluated strategies.

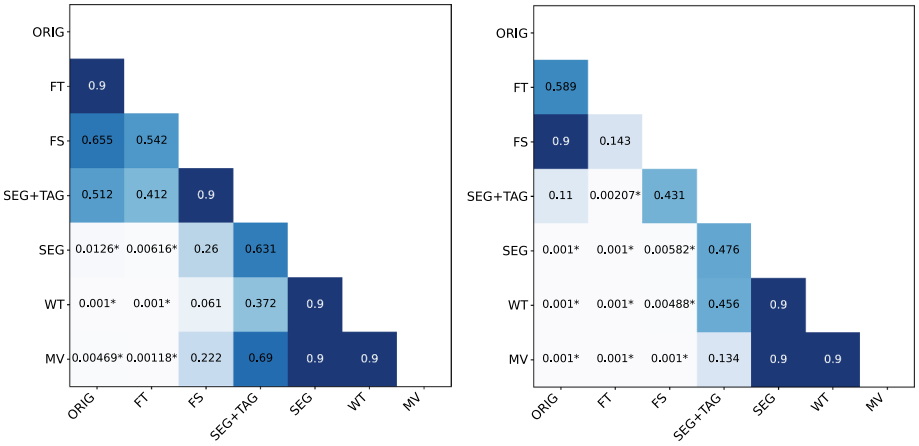
Table 2. ANOVA results across the evaluated configurations.

Metric	F	df_1	df_2	p-value	η^2
ROC	13.876	6	89.687	4.25×10^{-11}	0.481
F1-Score	7.137	6	90.049	3.04×10^{-6}	0.322
Recall	4.635	6	89.930	3.79×10^{-4}	0.236
Precision	3.163	6	90.023	7.33×10^{-3}	0.174

The results indicate that the null hypothesis can be rejected, suggesting that the different configurations do not perform equally. In particular, the treatment of the TAG has a significant effect on model performance, showing that its influence cannot be explained only by random variation.

6.3 Pairwise Comparisons: Which Marker Properties Influence Performance?

Once the global null hypothesis was rejected, we conducted pairwise comparisons using the Games-Howell test. Figure 4 shows the lower triangular matrices of p-values for the F1-score and ROC_AUC metrics.



(a) F1-score. (b) ROC_AUC.

Fig. 4. The p-values of the post hoc tests.

Presence Versus Absence of the Marker. Comparisons between configurations that preserve the TAG (ORIG, FT, FS, SEG+TAG) and those that remove or invalidate it (SEG, WT, MV) yield statistically significant differences ($p < 0.05$). By contrast, pairwise comparisons among ORIG, FT, FS, and SEG+TAG configurations rarely show statistically significant differences. This result indicates that preserving the marker is more important than maintaining its exact physical characteristics, reinforcing the idea that the TAG contributes functionally to the classification task.

Effect of Temperature and Marker Size. The lack of statistically significant differences between ORIG and FT in the metrics suggests that the specific thermal signature of the TAG is not a critical factor for the models' performance. Similarly, fixing the marker size does not result in systematic performance degradation.

These findings indicate that the model does not depend on an exact thermal pattern or a specific scale associated with the TAG, reducing concerns that the observed performance gains are due to trivial or non-physiological cues.

Effect of Marker Position: A Critical Factor. In contrast, the configurations in which the TAG was spatially displaced (MV) show statistically significant performance degradation when compared to the ORIG and FT configurations. Notably, the MV configuration often behaves similarly to the WT and SEG configurations in the pairwise analyses.

This pattern suggests that the spatial relationship between the marker and the breast anatomy plays a central role. When the marker is displaced, this relationship is lost, which reduces its usefulness as a spatial reference.

7 Conclusions

This study investigates the impact of non-physiological visual cues, particularly Anatomical Positioning Markers, on the training of deep learning models for breast thermogram classification. Based on the observation that segmentation and artifact removal approaches do not always result in performance improvements, we sought to understand whether such elements contribute functionally to the learning process.

The experimental results demonstrate that the marker cannot be interpreted as noise to be eliminated. Our analysis indicates that the way the TAG integrates into the visual context of the image influences how models organize and explore available spatial information. Particularly, spatial information was found to be more important than other physical properties, such as thermal signature or scale. Thus, the TAG's contribution relies on the geometric context of the image.

This work implies that pre-processing must be framed as a critical analytical step rather than a default filtering process. By underscoring the potential benefits of contextual elements for generalization, we extend the debate on bias, shortcut learning, and interpretability in standard benchmark datasets.

Future research aims to investigate the introduction of simplified and controlled artificial geometric references, seeking to assess whether such structural markers can replicate the effects observed with the original TAG. Future research also includes extending this analysis to other architectures and datasets to check whether these results remain consistent and to better understand how visual context affects the representations learned by the models.

Acknowledgments. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. The authors thank the financial support provided by National Council for Scientific and Technological Development – CNPq (313452/2025-3), Minas Gerais State Research Foundation – FAPEMIG, and Federal University of Juiz de Fora (UFJF). This work has been supported by High-Speed Integrated Research Network (RePesq).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Webb, R.: Digital revolution: how ai is reshaping science. Nature (2025)
2. Maitra, I.K., Bandyopadhyay, S.K.: Identification of abnormal masses in digital mammogram using statistical decision making. In: Hybrid Intelligence for Image Analysis and Understanding, pp. 339–368. Wiley (2017)
3. Gonçalves, C.B., Souza, J.R., Fernandes, H.: Cnn architecture optimization using bio-inspired algorithms for breast cancer detection in infrared images. Comput. Biol. Med. **142**, 105205 (2022)
4. Saadizadeh, S.: Breast cancer detection in thermal images using glrlm algorithm. In: International Conference on Electrical, Communication, and Computer Engineering (ICECCE), pp. 1–6. Kuala Lumpur, Malaysia (2021)
5. Bandyopadhyay, S.K.: Survey on segmentation methods for locating masses in a mammogram image. Int. J. Comput. Appl. **9**(11), 25–28 (2010)
6. Salmeri, M., Mencattini, A., Rabottino, G., Accattatis, A., Lojacono, R.: Assisted breast cancer diagnosis environment: a tool for dicom mammographic images analysis. In: IEEE International Workshop on Medical Measurements and Applications, pp. 160–165 (2009) IEEE
7. Gamal, S., Atef, H., Youssef, D., Ismail, T., El-Azab, J.: Early breast cancer screening from thermography via deep pre-trained edge detection with extreme gradient boosting. In: 2023 Intelligent Methods, Systems, and Applications (IMSA), pp. 430–433 (2023)
8. Fabbrizzi, S., Papadopoulos, S., Ntoutsi, E., Kompatsiaris, I.: A survey on bias in visual datasets. Comput. Vis. Image Underst. **223**, 103552 (2022)
9. Geirhos, R., et al.: Shortcut learning in deep neural networks. Nat. Mach. Intell. **2**(11), 665–673 (2020)
10. Silva, L.F., et al.: A new database for breast research with infrared image. J. Med. Imaging Health Inform. **4**(1), 92–100 (2014)
11. Nogales, A., Pérez-Lara, F., Morales, J., García-Tejedor, Á.J.: How do thermography colors influence breast cancer diagnosis? A hybrid model of convolutional networks with a weighted average evolutionary algorithm. In: Arai, K. (ed.) Intelligent Systems and Applications. IntelliSys 2022. LNNS, vol. 543, pp. 271–281. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-16078-3_17

12. Garyali, P., Ranjbar, I., Movahedi, S.: Screening of breast cancer from thermogram images by edge detection aided deep transfer learning model. In: 2022 IEEE 22nd International Conference on Bioinformatics and Bioengineering (BIBE), pp. 262–266. IEEE (2022)
13. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR) (2015)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015)
15. Zuluaga-Gomez, J.: A cnn-based methodology for breast cancer diagnosis using thermal images. Comput. Methods Biomech. Biomed. Eng.: Imaging Vis. **9**(2), 131–145 (2021)
16. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
17. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. J. Big Data **6**(1), 60 (2019)
18. Lu, Y., Chen, L., Li, Y., Kim, J., Wang, Z.: Data augmentation in classification and segmentation: a survey and new strategies. J. Imaging **9**(2), 46 (2023)