Breast Cancer Detection via Thermography Using Transfer Learning Architectures

Andrei Fernandes Zani
azani@alunos.utfpr.edu.br
Federal University of Technology – Paraná (UTFPR)
Apucarana, Paraná, Brazil

Daniel Prado Campos
danielcampos@utfpr.edu.br
Federal University of Technology – Paraná (UTFPR)
Apucarana, Paraná, Brazil

Kauane dos Santos Vieira
kauanevieira@alunos.utfpr.edu.br
Federal University of Technology – Paraná (UTFPR)
Apucarana, Paraná, Brazil

ABSTRACT

Breast cancer is the most lethal neoplasm among women in Brazil. Therefore, early detection is essential to initiate treatment promptly and reduce the risk of death. A public database of breast thermographic images was used for a classification task to distinguish between healthy and sick individuals. This process was conducted by a model based on transfer learning, in which a convolutional neural network pre-trained in a big dataset (ImageNet) is employed as a feature extractor and an artificial intelligence algorithm is trained on its top. To verify whether there are differences between different artificial intelligence architectures and approaches, two different convolutional neural network architectures were tested as feature extractors (VGG16 and ResNet), with two different approaches of artificial intelligence algorithms: the approach by neural networks and the approach by classical machine learning algorithms(Random Forest, Gradient Boosting, Support Vector Machine); with 8 models trained in total. In order to obtain statistically significant results, all models were trained in 5 different folds on cross-validation with 5 different seeds, yielding 25 distinct results for each model. The results indicated that the model that had the best performance in general was the approach that uses ResNet50 as an architecture with neural networks on top. Paired statistical tests, non-paired tests, and graphical analysis indicated that the models using ResNet50 as a feature extractor were superior to VGG16. Additionally, the models that used neural networks on top of the classifier had an advantage compared to those that used classical machine learning algorithms.

KEYWORDS

Thermography, Breast Cancer, Convolutional Neural Networks, Machine Learning

1 INTRODUCTION

This work explores the application of AI (Artificial Intelligence) techniques, specifically CNNs (Convolutional Neural Networks) and ML (Machine Learning), in breast thermograms, aiming for the early detection of breast cancer. The high mortality associated with the disease and the critical importance of diagnosis in the initial stages motivate the search for accessible and effective complementary screening methods. The infrared thermography, combined with CNNs and ML classifiers, presents itself as a promising tool in this context.

Breast Cancer represents a grave problem of global public health. It was the major cause of cancer related deaths between women worldwide and in Brazil, according to Bray et al. [1] and Instituto Nacional de Câncer [2]. Since early diagnosis is essential to improve patient prognosis [3], there is a great need for accessible diagnostic methods. While established imaging techniques such as mammography and ultrasound are widely used, they present limitations regarding discomfort, cost, radiation exposure, and reduced sensitivity in dense breast tissue [4]. In this context, infrared thermography has gained renewed interest as a non-invasive, radiation-free alternative capable of detecting early physiological changes [5]. This potential is significantly enhanced by recent advancements in AI, particularly through DL (Deep Learning) and ML, which are revolutionizing the analysis of breast thermograms and enabling the development of automated, precise diagnostic support systems [6].

The primary objective of this study is to evaluate the efficacy of a hybrid approach that combines pre-trained CNNs—specifically VGG16 and ResNet50 acting as feature extractors—with distinct ML classifiers (Neural Network (NN), SVM (Support Vector Machine), RF (Random Forest) and GB (Gradient Boosting)) to categorize breast thermograms as healthy or pathological. By identifying the optimal architecture-classifier combination, this research addresses the urgent need for improved early detection strategies amidst high breast cancer mortality rates. Capitalizing on the low-cost, non-invasive nature of thermography, the proposed integration with Artificial Intelligence aims to overcome the subjectivity of traditional visual interpretation. Consequently, this study seeks to validate a robust diagnostic support tool capable of facilitating large-scale screening and enhancing clinical decision-making, particularly in resource-constrained environments or for patients with dense breast tissue.

The remainder of this paper is organized as follows: Section 2 reviews related works and the state of the art; Section 3 details the methodology, including the dataset and the proposed architectures; Section 4 presents the experimental results and statistical analysis; and Section 5 outlines the conclusions and suggestions for future research.

2 RELATED WORKS

Recent advances in Artificial Intelligence applied to breast thermography have garnered significant attention due to the novelty and potential of this imaging modality. Current research largely focuses on Transfer Learning with CNNs and feature extraction techniques—ranging from deep features to handcrafted image descriptors like textures and edges—to train classical Machine Learning models. This section contextualizes recent developments and the methodologies employed in the field.

To demonstrate the generalizability of AI in medical diagnostics, Kermany et al. [7] employed Transfer Learning using the Inception

V3 architecture pre-trained on ImageNet. Although their work focused on optical coherence tomography and pediatric chest X-rays (for pneumonia detection), they achieved great performance across different pathologies. Their study validates the premise that CNN-based diagnosis is highly adaptable and applicable to a wide range of medical imaging tasks.

Focusing specifically on breast thermography, Aidossov et al. [8] compared the performance of several Transfer Learning architectures, including Xception, MobileNet, ResNet50, and VGG16. To mitigate the challenges of limited datasets and class imbalance, the authors employed data augmentation techniques (rotation, scaling, translation) and class weighting. Furthermore, they integrated a Bayesian network using clinical data (such as temperature extremes and thermal asymmetry) to enhance model interpretability. Their results indicated that Transfer Learning is effective even with small datasets, with the MobileNet architecture yielding the best performance among the tested models.

Alternatively, Hanieh et al. [9] proposed a "Deep Hybrid Network" approach that avoids pre-trained models. They designed a custom, lightweight CNN with four convolutional layers to act as a feature extractor. These extracted features were then fed into classical ML classifiers: SVM, K-Nearest Neighbors (KNN), and Fully Connected Networks. Their work demonstrates that it is feasible to achieve high diagnostic accuracy with reduced computational costs by using custom architectures instead of large pre-trained networks.

Finally, Youssef et al. [10] investigated a fusion strategy combining classical image processing with Deep Learning (DL). Their methodology involved extracting handcrafted features (using Gabor filters, Canny edge detection, and Histogram of Oriented Gradients) and merging them with deep features extracted from ResNet50 and MobileNet architectures. This combined feature vector was used to train SVM and XGBoost classifiers. The authors concluded that this hybrid approach, which integrates traditional feature extraction with deep representations, outperforms methods relying on either technique in isolation.

The reviewed literature demonstrates the potential of Transfer Learning and hybrid architectures for breast thermography analysis. While individual studies have explored specific CNN models or feature fusion techniques, there is a continuous need for a benchmarking of different feature extractors paired with distinct classification strategies. Motivated by these findings, this study aims to rigorously compare the efficacy of VGG16 and ResNet50 backbones when coupled with both NNs and classical ML algorithms (also called hybrid approach), as detailed in Section 3.

3 EXPERIMENTAL METHODOLOGY

The present section details the methodological procedures employed in this study. The workflow in Fig. 1 encompasses the dataset description, image pre-processing, and the application of Transfer Learning using established CNN architectures for deep feature extraction and ML classification. The next subsections will detail each of these methods.

3.1 Dataset

The current reasearch uses the public thermographic images developed and published by the Universidade Federal Fluminense in 2014[1] [11]. This dataset, at the present moment this article is being written, has more than 6500 infrared images from 330 patients, captured with static and dynamic acquisition protocols, classified as healthy when there is no tumor in the breast and cancer when there is. Regarding the dynamic and static images present in the dataset, only the static images were considered. Images that were blurry were also removed from the dataset. Therefore, the number of images was reduced to 1430, with 267 unique patients. The table 1 shows the image and patients distribution by class.

Class	Images	Patients	Proportion(%)
Class 0 (Healthy)	841	171	59
Class 1 (Cancer)	589	96	41
Total	**1430**	**267**	**100**

Table 1. Distribution of images and patients by class.

Table 1 reveals a slight imbalance between the healthy and cancer labels, with healthy images being overrepresented, which might cause a bias towards the majority class in the training of the models. The mitigation strategies and evaluation metrics adopted to address this issue are detailed in subsection 3.4. Furthermore, the images in this dataset were acquired using a FLIR SC620 thermal camera under controlled environmental conditions. The dataset includes records from multiple angles, including frontal, 45°, and 90° lateral views for both sides. Fig. 2 displays two samples from this dataset.

3.2 Image Pre-Processing

Before feeding the images to the pre-trained CNN, it is necessary to apply transformations to ensure they are in the standard format and compatible with the networks. Architectures such as VGG [12] and ResNet [13] specifically expect inputs at a resolution of 224×224 pixels with values normalized between 0 and 1. Consequently, we resized the images accordingly and normalized pixel values from the $[0, 255]$ range to $[0, 1]$. As noted by Goodfellow et al. [14], this normalization is a strictly mandatory step, as its absence prevents proper model convergence during training.

Furthermore, to mitigate the limited dataset size and enhance representativeness, data augmentation was employed to generate synthetic training samples. Following the protocols established by Aidossov et al. [8], these transformations included: (i) rotation, shifting by 0–10 degrees; (ii) scaling, randomly sampling the frame size between 80% and 110%; (iii) translation, shifting horizontally and vertically between -10% and 10%; and (iv) horizontal flipping, applied with a 50% probability.

3.3 CNN Deep Features and ML algorithms training

CNNs function as hierarchical feature extractors: initial layers detect simple patterns like edges, whereas deeper layers recognize

[1] Available in: https://visual.ic.uff.br/dmi

094

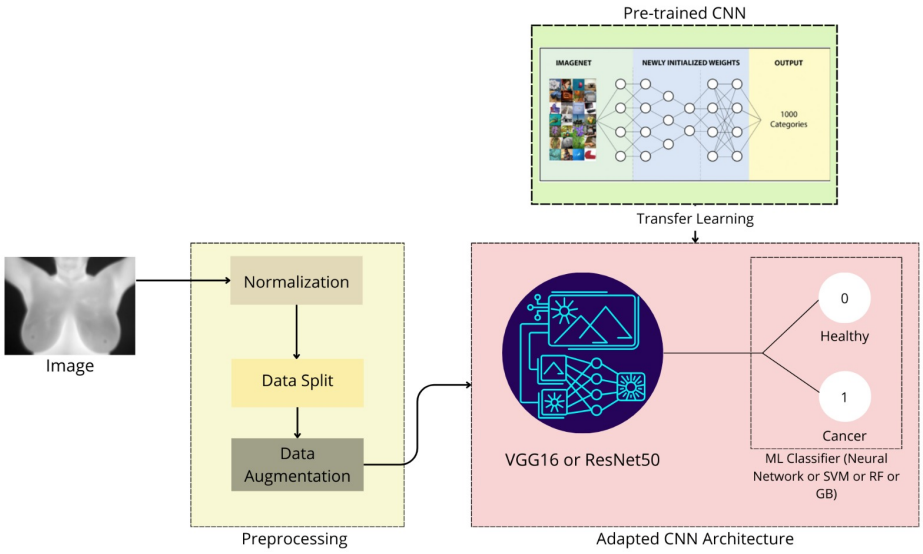

Figure 1. Schematic overview of the methodological steps. Transfer Learning image was adapted from Kermany et al. [7] and input image from da Silva et al. [11].

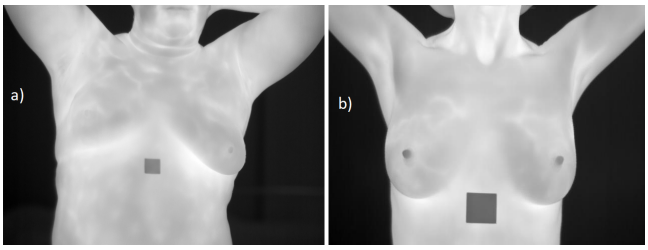

Figure 2. Comparison between healthy and cancer patients from the DMR-IR dataset [11]. (a) Shows a cancer patient; (b) Shows a healthy patient.

complex structures [15]. This work employs Transfer Learning strategies for breast thermography classification, using CNN architectures pre-trained on the massive ImageNet dataset to extract image features. By reusing pre-trained weights, the architecture serves as a feature extractor for both DL and ML models. This approach optimizes the binary classification task by eliminating the need to train a CNN from scratch, requiring only the adjustment of the architecture to suit the specific target task. For the task of extracting deep features, the selected architectures were VGG16 and ResNet50, due to their successful application in works such as Aidossov et al. [8], who employed both architectures, and Youssef et al. [10], who applied ResNet50 only.

To adapt the network to the target domain, the original classification head was discarded and replaced by different Machine Learning algorithms. The algorithms included a binary Neural Network output, SVM, RF, and GB. SVM and the NN were selected for their successful application in works such as Hanieh et al. [9]. Furthermore, the inclusion of RF and GB was motivated by the high performance of ensemble learning strategies, specifically boosting variants like XGBoost, as reported by Youssef et al. [10]. Additionally, the layers of the pre-trained architecture were frozen to prevent weight updates. This strategy preserves the learned features and mitigates the risk of overfitting inherent to training large-scale networks on smaller datasets.

To implement the NN in the classification head, the following architectural modifications were implemented: for VGG16, an additional dense layer of 32 neurons was inserted prior to the output stage. This addition was made to mirror its original design, which relies on fully connected layers to refine features [12]. In contrast, ResNet50 retained its standard structure as it uses Global Average Pooling, requiring no intermediate layers [13]. Regarding the classical ML algorithms (SVM, RF, and GB), no hyperparameter tuning was performed; default configurations were maintained to establish a baseline. The configuration details from the NN approach are presented in Table 2.

Due to memory limitations of the local hardware, a large batch size could cause out-of-memory errors and interrupt the process; therefore, the batch size was set to 16. The Early Stopping parameter refers to the training monitor employed: if the selected validation metric does not improve after a specific number of epochs, training is halted, and the model with the best performance is saved. Thus, the training was configured to stop if the validation F1 Score did not improve for 10 consecutive epochs. Given the class imbalance in the dataset, the F1 Score was selected as the monitoring metric,

095

Parameter	VGG16	ResNet50
Frozen Layers	Yes	Yes
Add. Dense Layer	32 units	None
Optimizer	Adam	Adam
Learning Rate	0.001	0.001
Loss Function	Binary Crossentropy	Binary Crossentropy
Batch Size	16	16
Max Epochs	30	30
Monitor (Early Stop)	val_f1_score	val_f1_score
Patience (Early Stop)	10	10

Table 2. Training configurations and hyperparameters per architecture.

as it effectively balances precision and sensitivity (recall) in such scenarios.

3.4 Experimental Evaluation

In order to evaluate the results and guarantee that there is no bias in the data that affects the output, experimental procedures were adopted. This subsection describes and details each one of them.

To ensure that results are not dependent on specific data splits or chance, and to analyze performance on unseen data, models were trained using five distinct random seeds and five cross-validation folds. Seeds are employed to introduce determinism into random operations—specifically defining data partitioning and neural network weight initialization—ensuring a specific, reproducible result for each run. In conjunction with this, K-fold cross-validation was applied to verify that the obtained performance was not an artifact of a specific data division. As described by Bishop [16], this technique divides the data into S equal groups, training on $S-1$ subsets and testing on the remaining one, repeating the cycle until all folds have served as the test set. In this study, 5-fold cross-validation was utilized (separating data into 4 parts for training and 1 for testing).

Crucially, specific sampling strategies were adopted to handle dataset characteristics. To mitigate the effects of class imbalance (59% healthy vs. 41% pathological, as shown in Table 1), stratified sampling was implemented. This ensures that the original class proportions are preserved in both training and testing sets, preventing bias due to disproportionate representation. Furthermore, since the dataset contains multiple thermograms per patient (as mentioned in subsection 3.2), a patient-wise grouping strategy was enforced to prevent data leakage. This approach guarantees that all images belonging to the same individual are kept exclusively within a single data split (either training or validation), preventing the model from being trained and validated on information from the same patient.

By combining 5 random seeds with 5-fold cross-validation, a total of 25 distinct results were obtained for each model. These 25 training runs are essential for analyzing the performance variability across different metrics and producing statistically significant results. These outcomes will subsequently be subjected to statistical tests to determine if there are genuine differences in performance between the evaluated models.

3.5 Statistical Tests

To determine whether the observed performance differences were statistically significant, two non-parametric hypothesis tests were applied to the F1-Score results. This metric was selected due to its suitability for imbalanced datasets, effectively balancing precision and sensitivity. Table 3 summarizes the statistical tests performed.

First, the Wilcoxon Signed-Rank Test was employed to evaluate the null hypothesis (H_0) that there is no difference in performance distribution between the VGG16 and ResNet50 feature extractors. As a paired test, it compares results obtained under identical experimental conditions; for instance, the F1-Score of "VGG16 + SVM" in a specific fold/seed is directly compared against "ResNet50 + SVM" in the same fold/seed. This procedure was repeated across all classifiers.

Subsequently, the Mann-Whitney U Test for independent samples was applied to compare the overall performance of the two methodological approaches: NN-based classification (DL) versus classical ML (also called Hybrid Approach). The null hypothesis (H_0) posited no statistical difference in the performance distributions of these two groups. Since this test is unpaired, it is appropriate for comparing independent sets with unequal sample sizes (e.g., aggregated scores from all DL models vs. aggregated scores from all ML models). The test effectively assesses whether a randomly selected observation from one group is likely to be greater than one from the other.

For both tests, a significance level of $\alpha = 0.05$ was adopted. The null hypothesis was rejected if the obtained p-value was less than α, indicating a statistically significant difference. Finally, significant differences ($p < 0.05$) were further analyzed regarding direction and magnitude through mean and median comparisons, as well as visual inspection using boxplots.

3.6 Evaluation Metrics

Performance metrics are used to evaluate how well the model handles the proposed task. The model outputs are binary: class 0 represents healthy images and class 1 represents cancerous images. Table 4 presents the Confusion Matrix structure regarding these predictions. The specific metrics used for evaluation are defined below, based on Marsland [17].

The models were evaluated using four performance metrics. Sensitivity (Recall), defined as $S = \frac{TP}{TP+FN}$, measures the proportion of positive instances correctly identified; this is critical in our study to minimize False Negatives and avoid misdiagnosing pathological cases. Precision, defined as $P = \frac{TP}{TP+FP}$, focuses on the reliability of positive predictions, which is essential to reduce patient distress caused by unnecessary follow-up exams. To balance these metrics on our imbalanced dataset, we employ the F1-Score ($F1 = 2 \cdot \frac{P \cdot R}{P+R}$), which penalizes disparities between the two scores [14]. Finally, the Receiver Operating Characteristic (ROC) curve illustrates classifier performance by plotting the True Positive Rate against the False Positive Rate ($1 - Specificity$), while the Area Under the Curve (AUC) provides a scalar measure of discriminative capacity [17].

3.7 Reproducibility

This study employed standard Data Science and Image Processing tools, primarily the Python programming language and its

096

Objective	Test	Null Hypothesis (H_0)	Data Input	Type
Compare Extractors	Wilcoxon	No difference in F1-Score distribution between VGG16 and ResNet50.	Scores paired by seed/fold (VGG16 vs. ResNet50).	Paired
Compare Approaches	Mann-Whitney	No difference in F1-Score distribution between DL and Hybrid approaches.	Aggregated scores grouped by approach (DL vs. Hybrid).	Unpaired

Table 3. Statistical tests, hypotheses, and data configurations.

	Predicted: Positive	Predicted: Negative
Actual: Positive	**True Positive (TP)** Correct positive prediction.	**False Negative (FN)** Incorrect negative prediction.
Actual: Negative	**False Positive (FP)** Incorrect positive prediction.	**True Negative (TN)** Correct negative prediction.

Table 4. Confusion Matrix Components Definition.

ecosystem. Model training and evaluation were conducted in a local development environment with the hardware specifications detailed in Table 5.

Component	Specification
Processor (CPU)	AMD Ryzen 5 5500 3.6GHz
Memory (RAM)	16 GB DDR4
Graphics Card (GPU)	NVIDIA GeForce RTX 2060 (6 GB VRAM)
OS	Windows 10 Pro 64-bit

Table 5. Experimental hardware specifications.

GPU acceleration was enabled using NVIDIA CUDA Toolkit 11.8 and cuDNN 8.6. The specific software libraries employed in this environment are listed in Table 6.

Library / Tool	Description
TensorFlow-Keras 2.10.1	CNN pre-trained models, preprocessing, and GPU training. (Selected as the last version with native Windows 10 GPU support).
NumPy	Mathematical operations and array manipulation.
Pandas	Analysis and manipulation of tabular data.
Matplotlib	Data visualization and plotting.
Scikit-learn	Loading and training of classical ML models.

Table 6. Libraries and development tools.

4 RESULTS

This section presents and discusses the results of the comparative evaluation of the two proposed approaches for breast thermographic image classification, comprising a total of 8 trained models. Following the methodological guidelines, model performance is primarily presented using boxplots, which summarize the distribution of metrics obtained during cross-validation across different seeds. The statistical significance of the observed differences is assessed using the non-parametric tests detailed in subsection 4.3 and subsection 4.4.

4.1 General Metrics Analysis

To better visualize the models' discriminative capacity, boxplots were generated for each relevant metric. Boxplots visually summarize the dataset distribution, displaying the median, interquartile range, and outliers. All charts are sorted from highest to lowest median. Complementing this visual analysis, Table 7 provides the mean and a 95% Confidence Interval, which are essential for comparing model metrics beyond visual estimation.

Among the four metrics analyzed, the ResNet50 + NN model performed best in Recall, F1-score, and AUC, achieving a lower score only compared to the Hybrid ResNet50 + SVM model in Precision. This indicates that the ResNet50 + NN model achieved the best generalization for the proposed problem, making it the most suitable approach for detecting tumors in breast thermographic images among the evaluated methods.

As illustrated in Fig. 3, the ResNet50 + NN model consistently achieved the highest recall across all folds. This visual trend is further supported by the metrics in Table 7, where the model reached a recall of 0.820 [0.797 - 0.844], effectively minimizing the false negative rate. Minimizing false negatives is critical, as a false negative result incorrectly indicates the absence of cancer when the neoplasm is actually present, posing a significant risk to the patient.

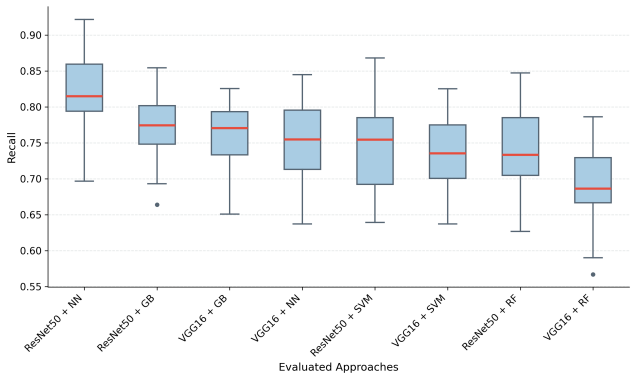

Figure 3. Comparative Recall scores across models.

Furthermore, the Hybrid ResNet50 + SVM approach outperformed the ResNet50 + NN model in mean Precision, ranking

097

CNN Model	Classifier	Performance Metrics (Mean [95% CI])			
		Precision	Recall	F1-Score	ROC AUC
ResNet50	NN	0.849 [0.829 - 0.870]	**0.820 [0.797 - 0.844]**	**0.833 [0.818 - 0.847]**	**0.857 [0.845 - 0.869]**
	GB	0.845 [0.828 - 0.862]	0.773 [0.753 - 0.793]	0.806 [0.792 - 0.820]	0.837 [0.825 - 0.848]
	RF	0.825 [0.805 - 0.845]	0.739 [0.713 - 0.766]	0.778 [0.760 - 0.797]	0.814 [0.799 - 0.829]
	SVM	**0.878 [0.859 - 0.897]**	0.751 [0.727 - 0.776]	0.808 [0.793 - 0.823]	0.839 [0.827 - 0.850]
VGG16	NN	0.842 [0.825 - 0.859]	0.753 [0.731 - 0.774]	0.793 [0.779 - 0.807]	0.826 [0.816 - 0.836]
	GB	0.765 [0.743 - 0.786]	0.758 [0.737 - 0.779]	0.760 [0.743 - 0.777]	0.798 [0.785 - 0.810]
	RF	0.792 [0.771 - 0.814]	0.691 [0.667 - 0.715]	0.736 [0.718 - 0.754]	0.781 [0.768 - 0.795]
	SVM	0.816 [0.797 - 0.835]	0.738 [0.717 - 0.759]	0.774 [0.758 - 0.791]	0.811 [0.798 - 0.824]

Table 7. Performance Comparison of the Models (Mean [95% CI])

among the top performers (Fig. 4). Although a slight overlap in their Confidence Intervals suggests comparable reliability, this metric measures the consistency of positive diagnoses. High precision minimizes the False Positive rate, making this model ideal for reducing unnecessary referrals of healthy patients for invasive or costly follow-up examinations, which can induce patient stress and anxiety. The performance gap between these two models in this metric, combined with their discrepancies in Recall and F1-Score, reveals that the SVM-based approach prioritized Precision at the expense of Recall during the inherent trade-off between these metrics. Despite its superior Precision, this significant compromise in Recall renders the ResNet50-SVM model less suitable as a primary diagnostic tool.

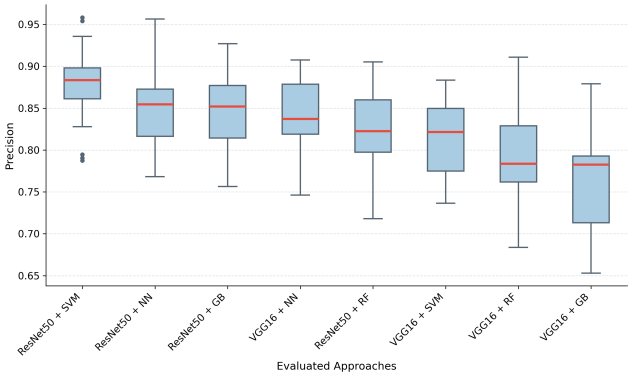

Figure 4. Comparative Precision scores across models.

The superiority of the ResNet50 + NN approach regarding the F1-Score demonstrates that it is the most suitable model for balancing Precision and Recall (Fig. 5). As discussed in subsection 3.6, the F1-Score metric penalizes models that disproportionately favor one metric over the other, which is a common issue when dealing with imbalanced datasets.

The high F1-Score indicates that this NN-based configuration successfully maintained a high Recall without generating an excessive number of False Positives. This implies that the model achieved the lowest rate of undiagnosed cancer patients without increasing the rate of erroneous positive diagnoses. This balance is ideal for medical screening applications because it identifies the maximum number of positive cases (high Recall), which prevents pathological patients from being incorrectly cleared while simultaneously minimizing false alarms (high Precision). Consequently, this reduces the referral of healthy patients for costly and stressful follow-up examinations and also confirms that the dataset imbalance did not significantly hinder the model's performance.

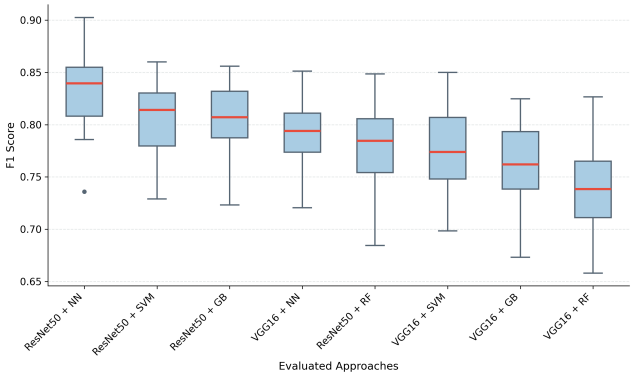

Figure 5. Comparative F1 Score across models.

4.2 ROC/AUC Analysis for Threshold Optimization

Finally, beyond threshold-dependent metrics, the ResNet50 + NN model achieved the highest AUC, confirming its reliability in distinguishing pathological from healthy patients regardless of the classification cut-off. The ROC curve was subsequently analyzed to identify an optimal decision threshold that maximizes sensitivity for screening purposes.

The optimal operating point was determined by calculating the minimum Euclidean distance to the ideal coordinate $(0, 1)$ (representing perfect sensitivity and specificity). Fig. 7 illustrates this optimized threshold of 0.27 (red point), which yielded a TPR (True Positive Rate) of 0.87 and an FPR (False Positive Rate) of 0.15. In contrast, the standard 0.5 threshold (blue square) resulted in a lower TPR of 0.73 with an FPR of 0.07. As detailed in Table 8, the optimized threshold improved both Recall and F1-Score. Although this

098

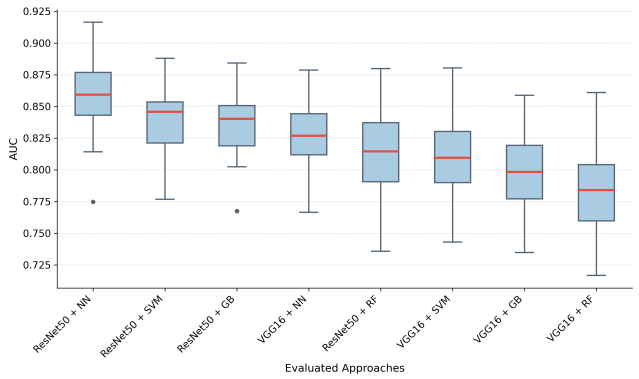

Figure 6. Comparative AUC Score across models.

adjustment incurred a slight reduction in Precision, the trade-off is clinically desirable; as stated in subsection 4.1, in medical diagnostics, prioritizing tumor detection (high Recall) takes precedence over minimizing false alarms, making the model more effective for the proposed application.

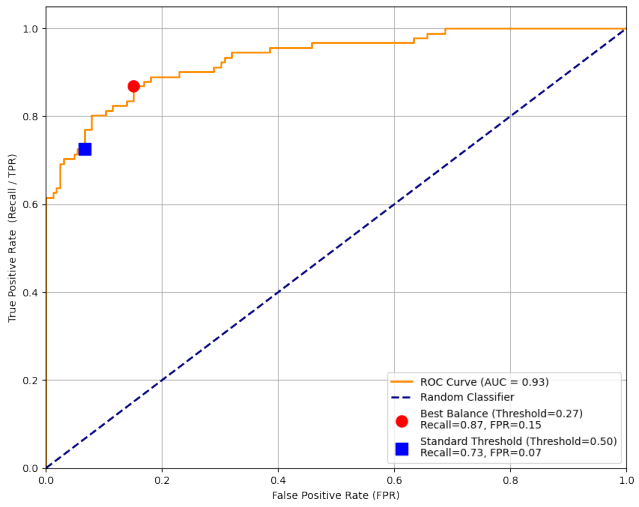

Figure 7. ROC Curve with the ideal threshold to improve the trade-off between recall and False Positive Rate.

Metric	Standard Threshold (0.5)	Optimized Threshold (0.27)
F1-Score	0.7857	**0.8103**
Precision	**0.8571**	0.7596
Recall	0.7252	**0.8681**

Table 8. Performance comparison by decision threshold.

4.3 Deep Feature Extractor Impact Analysis (VGG16 vs. ResNet50)

To isolate the impact of the feature extractor, a Wilcoxon Signed-Rank Test compared the F1-Scores of VGG16-based models against ResNet50-based ones, keeping the final classifier constant. The analysis yielded a statistically significant difference ($p < 0.001$), leading to the rejection of the null hypothesis. Fig. 8 illustrates this result, where the four asterisks (****) denote the highest level of significance ($p \leq 0.0001$). Visual inspection corroborates the statistical findings, revealing a distinct advantage for ResNet50-based approaches, which exhibit noticeably higher medians and lower dispersion compared to their VGG16 counterparts.

A plausible justification for this performance gap lies in the dimensionality of the feature embeddings extracted by each architecture. While VGG16 produces a 512-dimensional feature vector, ResNet50 yields a significantly larger output of 2048 features. This fourfold increase in dimensionality provides a richer representation of the input images, capturing details necessary to effectively distinguish between pathological and healthy patients.

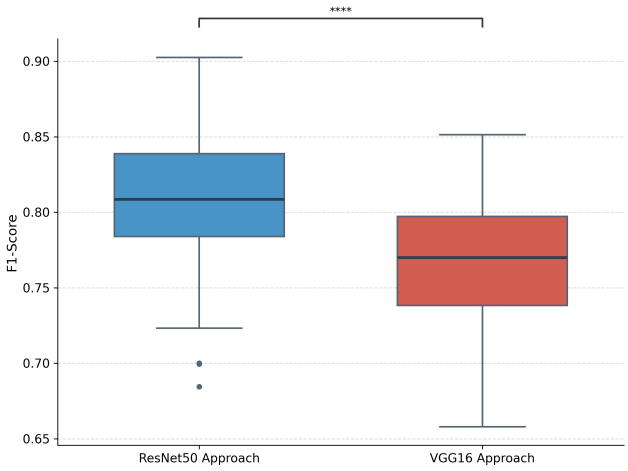

Figure 8. F1-Score comparative: ResNet50 vs. VGG16 Approach.

4.4 Classification Approach Impact Analysis (DL vs. Hybrid Approach)

A second analysis compared NNs (DL) against Classical Machine Learning (ML). Given the independent groups and unequal sample sizes, the Mann-Whitney U Test was employed, yielding a significant difference ($p < 0.001$). Fig. 9 visually reinforces this result, where DL approaches exhibit higher medians and lower dispersion compared to classical ML.

This performance gap is likely attributable to the regularization strategies employed. The NN models utilized techniques such as Early Stopping (Table 2) to actively prevent overfitting, whereas the classical ML models lacked equivalent dynamic regularization, potentially leading to overfitting. While this could be mitigated by performing a Grid Search to optimize hyperparameters [15], such an operation is computationally expensive. It requires training a new model for every parameter combination tested, significantly increasing the total computational cost compared to the efficiency of the approaches used in classical classifiers, potentially enhancing their generalization capabilities.

099

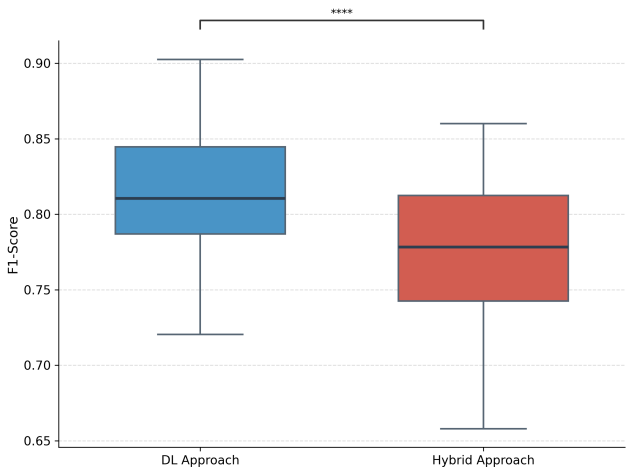

Figure 9. F1-Score comparative: DL vs. Hybrid Approach.

5 CONCLUSIONS

This article presented the development of breast thermographic image classification models using pre-trained CNNs and transfer learning, identifying the ResNet50 + NN approach as the top performer across most metrics. While the ResNet50 + SVM configuration achieved higher mean precision (despite interpolation between Confidence Intervals), the NN-based approach offered a superior balance between precision and recall, validated through a robust experimental protocol (5 folds, 5 seeds) that ensures statistical significance. Furthermore, statistical hypothesis testing conclusively demonstrated the superiority of ResNet50 over VGG16 as a feature extractor and confirmed that NN classifiers consistently outperform the Hybrid approach (with SVM, GB and RF) for this diagnostic task.

The primary challenges of this study involved developing a structured, reproducible codebase to manage the extensive volume of metrics and predictions generated across multiple training folds. Furthermore, constraints imposed by cloud-based environments necessitated the configuration of a local GPU setup. This migration required adjustments to the execution environment and the management of legacy library dependencies, which demanded significant configuration time but ensured the continuity of the experiments.

In future works, the focus will be on enhancing model performance and clinical applicability. It is intended to explore partial fine-tuning strategies for deep learning, alongside rigorous hyperparameter optimization (Grid Search) and dimensionality reduction for classical algorithms. To advance diagnostic utility, we will investigate multi-view analysis, patient-level evaluation, and interpretability using Grad-CAM. Finally, we will strengthen model generalization through external validation and by integrating different datasets, such as the one proposed by Rodriguez-Guerrero et al. [18].

In conclusion, the study successfully met its objectives, identifying the ResNet50 + NN architecture as the superior approach for breast thermography classification among the evaluated methods. The findings validate the feasibility of using these automated models as effective decision-support tools for healthcare professionals, demonstrating significant potential to streamline patient screening and assist in the early detection of breast cancer.

REFERENCES

[1] Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Isabelle Soerjomataram, and Ahmedin Jemal. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 74(3):229–263, 2024. doi: https://doi.org/10.3322/caac.21834. URL https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21834.

[2] Instituto Nacional de Câncer. *Controle do câncer de mama no Brasil: dados e números 2024.* INCA, Rio de Janeiro, 2024. URL http://www.inca.gov.br. Versão eletrônica.

[3] Ophira Ginsburg, Cheng-Har Yip, Aysha Brooks, Anne Cabanes, Marcello Caleffi, Jose Antonio Dunstan Yataco, Bishal Gyawali, Valerie McCormack, M McLaughlin de Anderson, Ravi Mehrotra, et al. Breast cancer early detection: A phased approach to implementation. *Cancer*, 126(Suppl 10):2379–2393, 2020. doi: 10.1002/cncr.32887.

[4] Sami Ekici and Hushang Jawzal. Breast cancer diagnosis using thermography and convolutional neural networks. *Medical Hypotheses*, 137:109542, 2020. ISSN 0306-9877. doi: https://doi.org/10.1016/j.mehy.2019.109542.

[5] Satish G. Kandlikar, Isaac Perez-Raya, Pruthvik A. Raghupathi, Jose-Luis Gonzalez-Hernandez, Donnette Dabydeen, Lori Medeiros, and Pradyumna Phatak. Infrared imaging technology for breast cancer detection – current status, protocols and new directions. *International Journal of Heat and Mass Transfer*, 108:2303–2320, 2017. ISSN 0017-9310. doi: https://doi.org/10.1016/j.ijheatmasstransfer.2017.01.086.

[6] Jose-Luis Gonzalez-Hernandez, Alyssa N. Recinella, Satish G. Kandlikar, Donnette Dabydeen, Lori Medeiros, and Pradyumna Phatak. Technology, application and potential of dynamic breast thermography for the detection of breast cancer. *International Journal of Heat and Mass Transfer*, 131:558–573, 2019. ISSN 0017-9310. doi: https://doi.org/10.1016/j.ijheatmasstransfer.2018.11.089. URL https://www.sciencedirect.com/science/article/pii/S001793101834033X.

[7] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018. doi: 10.1016/j.cell.2018.02.010. URL https://doi.org/10.1016/j.cell.2018.02.010.

[8] N. Aidossov, Vasilios Zarikas, Aigerim Mashekova, Michael Zhao, Eddie Ng, Anna Midlenko, and Olzhas Mukhmetov. Evaluation of integrated cnn, transfer learning, and bn with thermography for breast cancer detection. *Applied Sciences*, 13:600, 01 2023. doi: 10.3390/app13010600.

[9] Rezazadeh Hanieh, Saniei Elham, and Salehi Barough Mehdi. Enhancing breast cancer detection in thermographic images using deep hybrid networks. *Imaging and Radiation Research*, 7(1):6195, 2024. ISSN 2578-1618. doi: 10.24294/irr6195.

[10] Doaa Youssef, Hanan Atef, Shaimaa Gamal, Jala El-Azab, and Tawfik Ismail. Early breast cancer prediction using thermal images and hybrid feature extraction-based system. *IEEE Access*, 13:29327–29339, 2025. doi: 10.1109/ACCESS.2025.3541051.

[11] Lincoln da Silva, D. Saade, Giomar Sequeiros Olivera, Ari Silva, Anselmo Paiva, Renato Bravo, and Aura Conci. A new database for breast research with infrared image. *Journal of Medical Imaging and Health Informatics*, 4:92–100, 03 2014. doi: 10.1166/jmihi.2014.1226.

[12] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. URL https://arxiv.org/abs/1409.1556.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

[14] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, 2016.

[15] Vijay Janapa Reddi. *MLSysBook.AI: Principles and Practices of Machine Learning Systems Engineering*. 2024. URL https://mlsysbook.org. Available at: https://mlsysbook.org.

[16] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

[17] Stephen Marsland. *Machine Learning. An Algorithmic Perspective 2nd ed.* CRC, 2015. ISBN 9781466583337; 1466583339.

[18] Steve Rodriguez-Guerrero, Humberto Loaiza-Correa, Andrés-David Restrepo-Girón, Luis Alberto Reyes, Luis Alberto Olave, Saul Diaz, and Robinson Pacheco. Dataset of breast thermography images for the detection of benign and malignant masses. *Data in Brief*, 54:110503, 2024. ISSN 2352-3409. URL https://www.sciencedirect.com/science/article/pii/S2352340924004724.

100