10.21611/qirt.2020.100

QIRT PORTO – PORTUGAL 2020, 6-10 JULY

15th Quantitative InfraRed Thermography Conference

Automatic breast cancer detection on breast thermograms

By R. Gonzalez-Leal*, M. Kurban*, L.D. López-Sánchez*, and F.J. Gonzalez*

* Eva Tech, Diagonal Patriotismo 12 P-3 Col. Hipódromo Condesa, Ciudad de México 06100, México.

raymundo.gonzalez@evacenter.com, kurbanrita@gmail.com, luis.lopez@evacenter.com, javier.gonzalez@evacenter.com

Abstract

Widespread use of thermography as a method for breast cancer detection has been limited due to the lack of standard interpretation methods, relying mostly on subjective analysis. Some studies address this issue by proposing quantitative approaches. Such automated assessment methods have led to promising results, yet these methods have received limited attention. In this study, we propose a computerized system for the interpretation of breast thermograms. Our system consists of an automatic breast segmentation step, an acquisition device-dependent image processing pipeline, and an automated feature extraction pipeline. These features include expert-designed evaluation methods as well as texture and statistical features.

1. Introduction

Breast cancer is prevalent all around the world. Despite a high success rate of modern treatments with 95% survival upon early detection, the disease remains one of the leading causes of death in women due to the difficulty of accurately identifying early-stage tumors [10]. Thus, early detection is the key to reducing breast cancer mortality rate. Mammography has long been considered the gold standard for breast cancer detection. While it is currently the best screening method available, its performance is affected by breast density. The potential of infrared thermography as an alternative breast cancer detection tool has been studied for decades. Thermography is a safe and tissue-agnostic method [8] that can close the gap in the early prevention of breast cancer in young women and women with dense breasts. Thermography methods have long suffered from inaccurate equipment and a lack of objective evaluation methods, which decreased their popularity in the medical community.

The Thermal Score approach, cited in Gonzalez [7], is the most well-known effort to introduce quantitative guidelines for breast thermography interpretation. The score is defined as the sum of the vascularity value and a ΔT value. The ΔT value is measured as the temperature difference, in degrees Celsius, at the lesion site compared to the contralateral breast. The automation of the score makes it possible to quantify human performance and compare it to that of the Computer-Aided Diagnosis (CAD) systems for thermal imaging. CADs represent a complementary tool that could increase thermal imaging usage as a diagnostic and screening technique for breast cancer detection.

In this study, we introduce a fully automated CAD that streamlines region of interest (ROI) segmentation, image processing, feature extraction, and automatic evaluation of explorations of multiple-source and multi-protocol breast thermograms that reach state-of-the-art performance [3,22]. Furthermore, this study is unique in the diversity and size of its data source.

2. Data

Data was collected from diverse sources, including the Visual Lab DMR database [20], previous research datasets collected by one of the authors and used in Morales et al. [15] and Gonzalez [7], and the non-profit organizations ASBIS and Manos Rosas in Mexico. The thermographic image database consisted of three views, one frontal and two laterals.

To ensure our results' replicability and generalizability, we discarded data that could confound or bias our model based on the following criteria:

- Patients without a screening study — mammography for women after 40 and ultrasound for younger patients.
- Patients with a BI-RADS 0 in their screening study unless a follow-up result was available.
- Patients with mastectomy.
- For patients with multiple thermograms, we kept only the most recent exam, except for exams performed after the beginning of breast cancer treatment.
- BI-RADS 3 (probably benign) was also excluded from the study. Unlike BI-RADS 1 and 2—normal and benign—or BI-RADS 4 and 5—suspicious or highly suspicious—, BI-RADS 3 has different meanings for different screening modalities as mammograms and ultrasound [12].

The resulting sample consisted of 78 cancer patients, 34 patients with abnormal findings but no breast cancer, 1657 patients with BI-RADS 2, and 24 patients with BI-RADS 1. The mean population age is 43.3 years, with an STD of 12.7. The dataset was divided into a training set of 1187 patients and a test set of 606 patients using a random sampling method. The test data was not used to make modeling decisions to avoid getting overly optimistic outcomes that do not generalize well.

3. Methods

3.1. Segmentation

The first step towards the automatization of thermography interpretation is breast segmentation. Segmentation refers to identifying relevant sections of thermal images that correspond to each breast, axillary region, and nipple area. By performing segmentation, we ensure that irrelevant segments of the image, such as the abdomen and background, do not impact our analysis. The pixels outside of the region of interest can be considered as sources of noise. By excluding them from the study, we are cleaning the signal so that our machine learning models require fewer observations to achieve decent performance.

We use the MXNet ResNet34 neural network, pre-trained with Imagenet. Residual neural networks such as this one can achieve higher performance than traditional convolutional neural networks since they can be deeper without exhibiting the vanishing gradient problem. We have fine-tuned the neural network using the ground truth values obtained through the AWS Sagemaker's labeling jobs. Fine-tuning was our strategy of choice since features from the first layers of the model recognize simple textures and patterns typical for a variety of images, including thermograms. Hence, we obtained results by freezing the first layers of the model and fine-tuning the few thousand parameters in the last layers. Custom data augmentation policies were used to enhance the fine-tuning results.

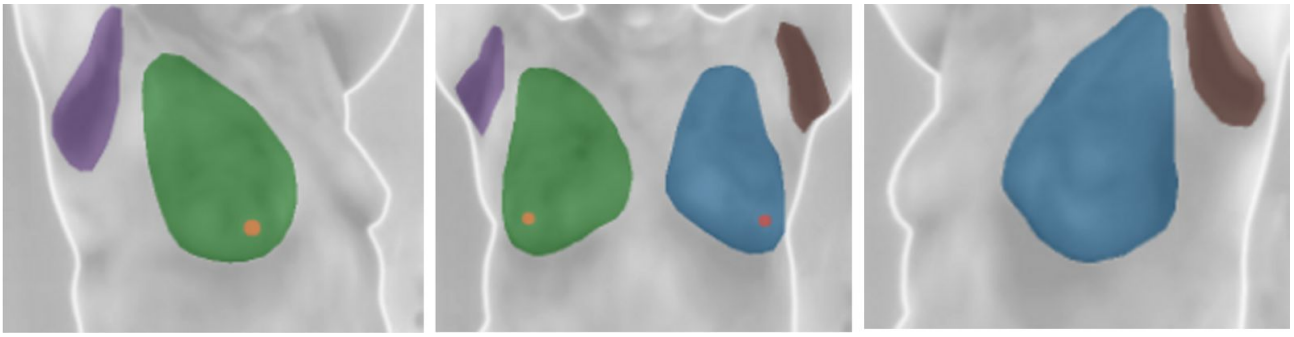

Figure 1. *The outcome of the breast segmentation model. Colors indicate segmented areas: breasts, nipples, and armpits. From left to right: right, front, and left views of the patient.*

3.2. Contrast Enhancement and CLAHE

After performing segmentation, we obtain the bounding box for each breast by iterating through the segmentation output and finding the minimum and maximum coordinate values. Pixels inside the bounding box that do not correspond to the region of interest are turned to null values, ensuring that we exclude them from the analysis. After that, we resize the matrices to 96x96 pixels to standardize input dimensions.

Each of the temperature matrices obtained through the breast bounding boxes goes through preprocessing consisting of a Gaussian filter and an application of the CLAHE algorithm. The Gaussian filter decreases input noise, while the CLAHE algorithm highlights local temperature differences, which improves the correlation of features with the BI-RADS status. The combination of these elements has also been found to contribute to domain adaptation. In research settings in which data comes from different sources, this preprocessing homogenizes the input reducing source-specific differences. This process prevents clustering by the data source in t-distributed Stochastic Neighbor Embedding (t-SNE) plots, which, as explained later, is an informative validity test. Final preprocessed explorations consist of four images: lateral and frontal thermograms of right and left breasts:

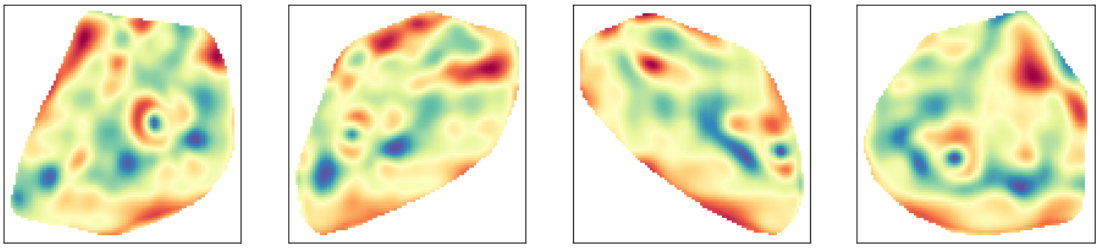

Figure 2. *The visualization of pre-processed images with a Gaussian filter after the application of the CLAHE algorithm. From left to right: lateral and frontal images of the right breast, frontal, and lateral images of the left breast.*

3.3. Features

We developed algorithmic Delta-T and vascular asymmetry scores to formalize expert systems and fed them into machine learning algorithms along with texture descriptors such as statistical features, Haralick features, Local Binary Patterns, and Histogram of Oriented Gradients.

3.3.1. Statistical

The main idea behind the statistical features is to extract summary statistics to represent the temperatures of each breast and compare breasts through these statistics. We calculated each statistic from the list below for each breast and stored the absolute difference:

1. Entropy. The average level of uncertainty inherent in the random variable's possible outcomes. We treat a temperature histogram for each breast as the distribution of a random variable. Breasts with less uniform thermal patterns have a higher entropy value.
2. Kurtosis. The fourth normalized moment, which is a measure of how likely the distribution produces outliers.
3. Difference in maximums. The difference between the maximum temperature of each breast. High values indicate a region with an abnormally high temperature in one of the breasts.
4. Difference in minimums. We compute the difference between the minimum temperature value of each breast. High values indicate that one of the breasts is significantly hotter than the other.
5. Standard Deviation: A measure of the amount of variation in the temperature set. A higher deviation signals less uniformity in the breast, which may be caused by regions of abnormality.
6. Skewness. The asymmetry of a distribution around its mean. High skewness in a breast can be an indicator of points of hyperthermia or hypothermia skewing the temperature distribution towards the right or the left end, respectively.

3.3.2. GLCM

The Gray Level Co-occurrence Matrix (GLCM) is a matrix that holds the probabilities of spatial relationship occurrences between gray level pixels within an image. Haralick [9] proposed a set of features extracted from the GLCMs yielding characterization of textures as smooth, coarse, grainy, etc. They are now widely used in breast cancer detection literature. For example, Acharya et al. [2] used GLCM features and SVM on a balanced set of 50 thermograms and achieved the sensitivity and specificity scores of 85.71% and 90.48%, respectively. Another peer-reviewed article [14] reported a sensitivity of 78.6% using a k-NN classifier and 20 GLCM features. Thus, despite limited datasets and, in some cases, unclear methodology, these papers offer a promise for the use of GLCMs for breast cancer detection.

We compute 2 GLCMs per temperature matrix, using a fixed offset and different angles so that each temperature matrix has a GLCM calculated in the horizontal and vertical directions. For each Haralick function, we obtain the feature's average for each breast. We average across 4 GLCMs to get a feature value for the breast since we compute two GLCMs per thermal matrix and store two thermal matrices per breast —a frontal and a lateral view. For each feature, we store the comparison of the feature value between breasts, either absolute values or actual differences.

3.3.3. HOG

The logic behind the Histogram of Oriented Gradients (HOG) is that the distribution of intensity gradients can represent shapes within an image. Abnormal shapes and dramatic shifts in intensity can be signs of pathological changes in breast tissue and can be useful for breast cancer detection.

Raghavendra [18] extracted HOGs from a balanced set of 50 patients and demonstrated that this approach, combined with dimensionality reduction techniques, leads to a significant improvement in several metrics compared to other feature extraction approaches. Ergin [5] obtained similar findings and proposed a HOG-based computer-aided diagnosis framework to aid radiologists.

We partially replicated the approach used in Raghavendra [19] and improved the method by experimenting with different dimensionality reduction techniques. Figure 3 presents a visualization of this feature vector. Blue color indicates a homogeneous surface, while green and yellow indicate a significant change in image intensity. Similarly, rotated shapes suggest that there is a change in the gradient direction:

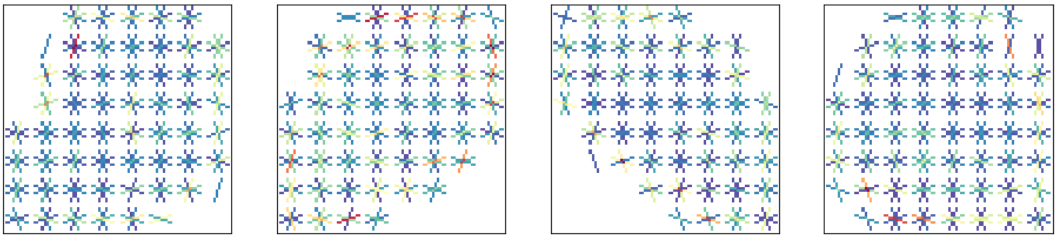

Figure 3 . *The visualization of HOGs. From left to right: lateral and frontal images of the right breast, frontal and lateral images of the left breast.*

The HOG feature descriptor yields highly-dimensional vectors. Processing such data for thousands of patients is computationally intensive. Additionally, these vectors have redundant features that add no relevant information because they are correlated or only contain zeros. Thus, we compared three techniques — Kernel Principal Component Analysis (KPCA), Independent Component Analysis (ICA), and Locality Preserving Projections (LPP). We found that KPCA consistently outperformed other methods independent of the kernel, number of components, and the train-validation split. Thus, we picked this approach to produce the final vector, which consists of X features that contain information on essential components of the original vector.

3.3.4. LBP

Local Binary Patterns (LBPs) are a texture descriptor popularized by Ojala [16]. The value of this feature vector in cancer detection is that it helps to identify the hottest areas of the breast that might indicate a pathological increase in the temperature. It also highlights areas of increased vascularity that are indicative of abnormality.

This method is widespread in breast cancer detection literature and has been used both with thermographic [1] and mammographic [18] imagery. It has also been applied for vascular area segmentation and classification [1,13,18].

Choi et al. [4] classified 303 mammograms as either breast masses or healthy tissue, using the rotation-invariant LBPs approach. The reported AUC score was around 90%. Li et al. [13] proposed a method for mass segmentation and detection using a rule-based algorithm called MCL (multiple concentric layers). They trained the model on 125 images and achieved the average recall of 76.8%. The main criticism of this approach is that MCL is empirically optimized and might not generalize well. Another study [1] reports the AUC score of 99% on a sample of 50 patients used for training, validation, and testing. Though we might not obtain the same performance on a more massive dataset, these findings make a strong case for the use of LBPs for automatic detection.

The idea behind the LBP operator is quite simple: it replaces pixels of the image with numbers that encode the local structure around each pixel. Specifically, we compare each central pixel with its eight neighbors. If the neighboring pixel's value is smaller than that of the central pixel, it gets a value of 0. If the value is equal to or greater than that of the central pixel, it receives the value of 1. After that, we generate a binary number for each of the central pixels by concatenating these binary bits. The resulting decimal value replaces the original central pixel value.

The LBP value of 0 corresponds to the hottest area of the breast because it means that the pixel does not have any hotter neighbors. Pathological changes in the tissue might cause such a localized increase in temperature. The figure below shows how this feature vector highlights the vascular patterns of the breast:

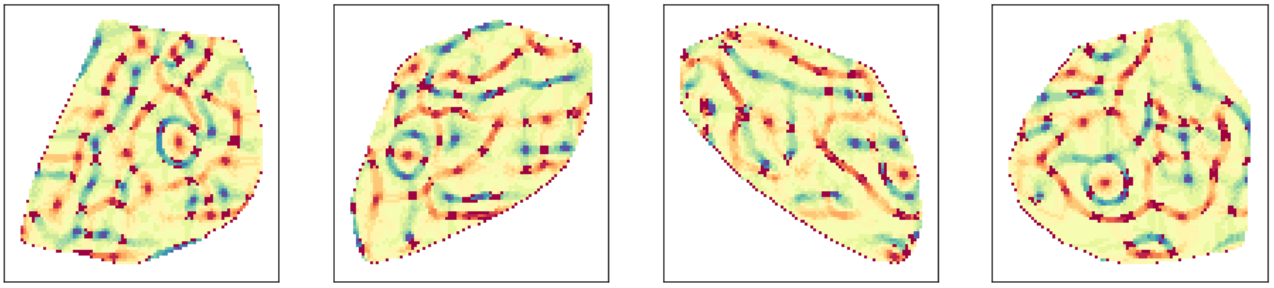

Figure 4. *The visualization of LBPs. From left to right: lateral and frontal images of the right breast, frontal and lateral images of the left breast.*

The resulting feature vector is used in the machine learning pipeline. It consists of 20 features that represent the difference in the LBP patterns of the left and right breasts after applying the kernel PCA.

3.3.5. Automatic Thermal Score Extraction

ΔT measurements were computed by automatically locating high vascularity areas and comparing its temperature to that of the counter-lateral site.

We automated the thermal score presented in Gonzalez [7] using the LBPs and the random walker segmentation. The LBPs reveal the local maxima inside the breast (LBP value of 0), which are the points at which ΔT is calculated. In the Random Walker algorithm, we mark seed regions: areas of the highest and lowest temperatures. For every unlabelled pixel, the algorithm initializes a random walker that can go anywhere in the image. Then, it calculates the probability that each random walker reaches one of the labeled pixels. By assigning each pixel to the class with the

highest probability, obtaining a segmentation of good quality is possible. This approach allows us to quantify the extension of vascular patterns, and whether vascularity is distributed similarly in each breast.

We combined two components of the score and experimentally identified that the correlation between the automated thermal score and the human evaluations is equivalent to the correlation between two interpreters on a set of 114 manually labeled thermograms. Hence, we have developed powerful tools capable of automatically producing results similar to those that well-trained interpreters.

4. Validity Checks

The literature on breast cancer detection using machine learning and infrared thermography does not have a scarcity of publications reporting suspiciously high classification performance. No matter how good the classification system is, it will not be able to overcome the physical limitations of the technology. The thermal signal a tumor produces on the breast surface decreases as tumor size and aggressiveness decrease, and as tumor depth increases. It has been estimated that a thermal image with a sensitivity of 20mK will be able to detect tumors with 3cm in diameter at most at 7cm below the skin, and tumors with 0.5cm in diameter at no more than 2.5 cm below the breast surface [6]. Hence, results that are near-perfect performance likely suffer from validity limitations that undermine their generalizability. There are two sources of bias that may undermine the validity of machine learning results: the data and the model evaluation procedure. Fortunately, one can prevent and test for sources of bias.

4.1. T-SNE Plots

Heterogeneous data sources can become a dangerous source of bias. Some data sources have a much higher proportion of cancer cases, different demographics, or other factors that can confound results. Thus, we ensure that preprocessing steps homogenize images and prevent models from learning biased representations. We use T-distributed Stochastic Neighbor Embedding (t-SNE) to ensure that multiple centers do not introduce bias to the model. Our test consists of creating t-SNE plots for various parameter values and determining whether exams cluster by some factor that is not the cancer status. We represent positive and negative cases with different shapes, and color data points based on the potential confounder factor. A problematic t-SNE plot would reveal clusters by such factors. For example, clusters formed by the center would mean that the model can discern the origin of the data based on the extracted features. However, this is not the case for our data. A combination of preprocessing techniques described above enabled us to homogenize the data and make sure there is no clustering by confounding variables, as illustrated by the plot below. The distribution is similar for multiple parameter values of the t-SNE plot.

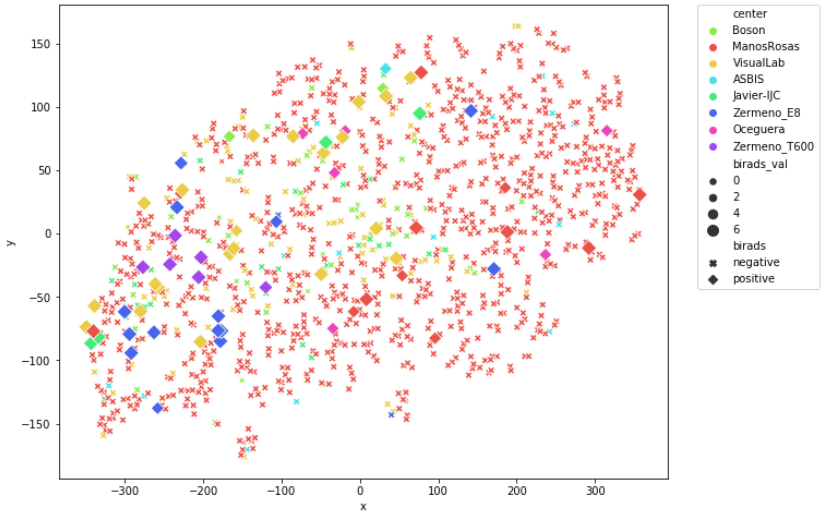

Figure 5. *A t-SNE plot of the training set explorations. Colors represent different centers. The size of the marker indicates BI-RADS status. The shape of the pointer indicates the status of the exploration.*

4.2. Data Leakage

Data leakage occurs when information from outside the training set is used to influence model decisions. Data leakage is common and even occurs in peer-reviewed articles. For example, researchers have claimed both sensitivity and specificity of 100%:

Classifiers	Sensibility	Specificity	Precision	Accuracy	ROC area
Bayes Net	100.00%	100.00%	100.00%	100.00%	1.00
Multi-layer perceptron	92.59%	96.30%	96.15%	94.44%	0.96
Decision table	92.59%	92.59%	92.59%	92.59%	0.92
Random forest	96.30%	92.59%	92.86%	94.44%	0.99
Average	95.37%	95.37%	95.4%	95.38%	0.97

Classification confusion matrix with Bayes Net	With cancer	Healthy
Classified as with cancer	40	0
Classified as healthy	0	40

Table 1. *Results table from a published article that presents biased performance due to the absence of a separate test set [21].*

However, the validity and generalizability of such results are highly questionable. For example, the authors of the above-mentioned peer-reviewed article used cross-validation to make model decisions and evaluate models. Although they used a different type of cross-validation, namely LOOCV, as their "test," it does not guarantee the absence of data leakage. One may think that changing the number of folds is enough to argue that no model decisions were made with the evaluation data since the evaluation was run with a different number of folds. To demonstrate that this approach is faulty, we ran a set of tests where we randomly reassigned the training and test set labels without changing the training and evaluation process. Since there is no relationship between inputs and target data are now randomized, the model should not perform above random chance. Performance significantly above chance would necessarily mean that the model evaluation method is flawed.

We replicated the approach described above using subsets of our data. The average area under the ROC curve was close to 0.9, and the most common outcome was 1.0, which corresponds to perfect performance. Thus, performing model selection and evaluation on the same set leads to perfect performance with completely random labels, which indicates that their results cannot be trusted.

Unlike the approach described above, we perform the train-test split before making any modeling decisions and only use our held-out test set for final model evaluation. Running the same test using this alternative approach results in the AUC score of 0.5, which is the expected value for a valid method. Hence, the train/test split approach, unlike CV with the different numbers of folds, is not prone to bias. Since the clustering analysis revealed no bias in the data and our overfitting test reveals that the process we follow does not bias results, we believe that our scores are generalizable.

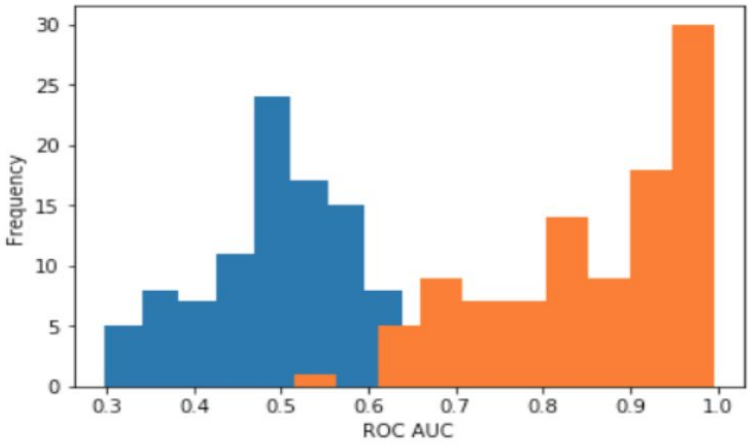

Figure 6. *The distribution of AUC-ROC scores after multiple runs with the biased approach (orange) vs. train-test split approach (blue).*

Another opportunity for data leakage arises when researchers and companies claim high performance but do not disclose their data obtention and model selection techniques [15]. As a result, such models cannot be put through the validity tests described above. Thus, results cannot be replicated, and their generalizability is questionable.

3. Results

We constructed a binary target variable by setting BI-RADS 4-6 to positive and BI-RADS 1-2 to negative. We used the Area Under the Curve for the Receiver Operating Characteristics curve (AUC ROC) metric to evaluate model

performance. AUC ROC is an intuitive metric that illustrates the best trade-off between sensitivity and specificity, often reported in the thermography literature.

We trained a Logistic Regression model and tuned hyperparameters using 5-fold cross-validation on the training set. The held-out test set of 606 patients was never used to make any modeling decisions and was only used for final evaluation. The figure below shows the ROC curve for the model, as well as the outcomes of the overfitting test presented above. The model obtained an area under the ROC curve of 0.785. The overfitting test resulted in an AUC ROC of about 0.5 on a test set with randomly generated labels, hence revealing no signs of data leakage and suggesting the generalizability of our results.

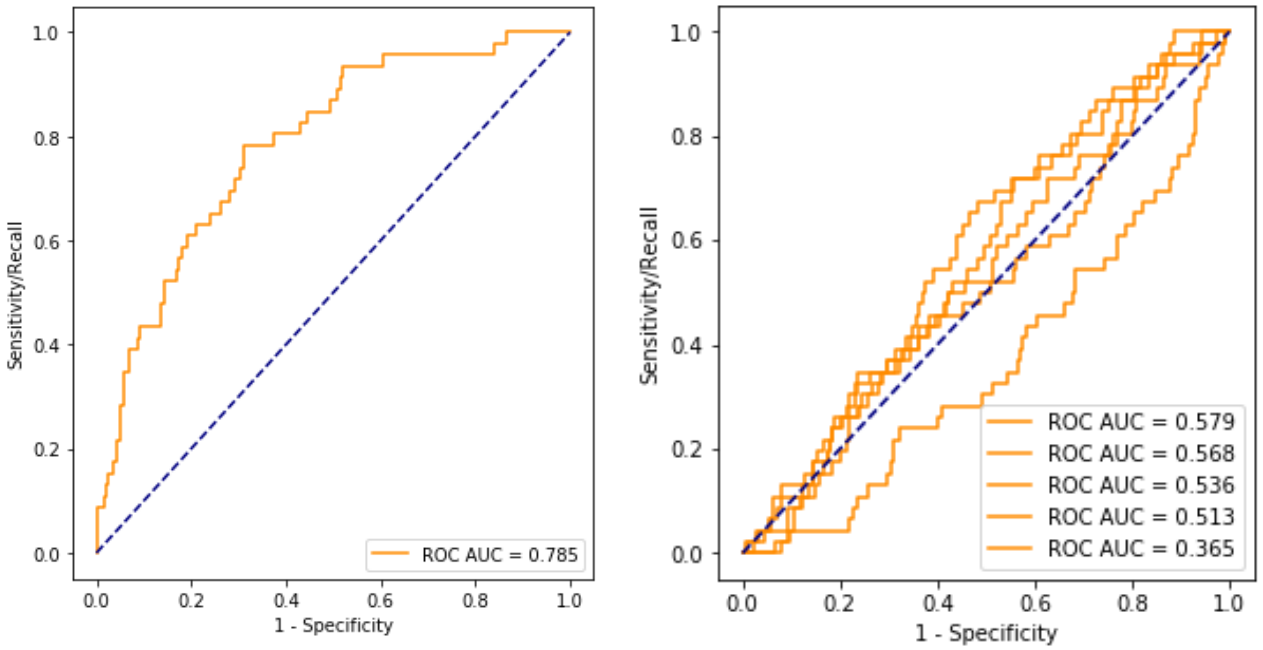

Fig. 8. *Left: ROC-AUC curve for logistic regression. Right: Overfitting test with random labels.*

4. Conclusion

Thermography for breast cancer detection is a widely studied approach that has received mixed evaluations. Some advantages of this method include reduced costs and improved screening accessibility and the mitigation of potential risks associated with other detection methods. Thermography can be performed more frequently, which could lead to earlier cancer detection in some contexts, improving patient survival rates.

This paper advances the thermography literature by offering a novel perspective on the creation and evaluation of machine learning methods for breast cancer detection. We created a fully automated CAD system that extracts regions of interest, preprocesses the images to ensure homogeneity, extracts features from processed images, and automatically evaluates breast thermograms. Throughout the paper, we highlight the importance of high research standards and validity checks. We argue that in order to trust ML results, researchers need to have clear selection criteria, avoidance of confounders and bias in the data, and model selection and evaluation processes. We designed a transparent and reproducible pipeline that achieved performance comparable with state of the art on an imbalanced set of images, and that proved robust to the validity checks proposed in this work.

REFERENCES

[1] Abdel-Nasser, Mohamed, Antonio Moreno, and Domenec Puig. "Breast cancer detection in thermal infrared images using representation learning and texture analysis methods." Electronics 8.1 (2019): 100.
[2] Acharya, U. R., Ng, E. Y. K., Tan, J.-H., & Sree, S. V. (2010). Thermography Based Breast Cancer Detection Using Texture Features and Support Vector Machine. Journal of Medical Systems, 36(3), 1503–1510. doi:10.1007/s10916-010-9611-z
[3] Borchartt, T.B.; Conci, A; Lima, R.C.F; Resmini, R.; Sanchez, A. "Breast thermography from an image processing viewpoint: A survey". Signal Processing 9(10) (2013): 2785-2803.
[4] Choi, Jae Young, Dae Hoe Kim, and Yong Man Ro. "Combining multiresolution local binary pattern texture analysis and variable selection strategy applied to computer-aided detection of breast masses on mammograms." Proceedings of 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics. IEEE, 2012.
[5] Ergin, Semih, and Onur Kilinc. "A new feature extraction framework based on wavelets for breast cancer diagnosis." Computers in biology and medicine 51 (2014): 171-182.
[6] Gonzalez, Francisco. (2007). Infrared imager requirements for breast cancer detection
[7] Gonzalez, Francisco. (2011). Non-invasive estimation of the metabolic heat production of breast
[8] González, F. J., Ríos, J., González, R., & Cruz, O. (2020). Effect of tissue density on the temperature pattern of the breast. In Medical Imaging 2020: Physics of Medical Imaging (Vol. 11312, p. 113125J). International Society for Optics and Photonics.
[9] Haralick, Robert M. "Statistical and structural approaches to texture." Proceedings of the IEEE 67.5 (1979): 786-804.
[10] Kennedy, Deborah A., Tanya Lee, and Dugald Seely. "A comparative review of thermography as a breast cancer screening technique." Integrative cancer therapies 8.1 (2009): 9-16.
[11] Keyserlingk, John R., et al. "Functional infrared imaging of the breast." IEEE Engineering in Medicine and Biology Magazine 19.3 (2000): 30-41.
[12] Lee, K. A., Talati, N., Oudsema, R., Steinberger, S., & Margolies, L. R. (2018). BI-RADS 3: current and future use of probably benign. Current radiology reports, 6(2), 5.
[13] Li, C. I., D. J. Uribe, and J. R. Daling. "Clinical characteristics of different histologic types of breast cancer." British journal of cancer 93.9 (2005): 1046-1052.
[14] Milosevic, M., Jankovic, D., & Peulic, A. (2014). Thermography based breast cancer detection using texture features and minimum variance quantization. EXCLI journal, 13, 1204
[15] Morales, Antony & Kolosovas, E. & Guevara, Edgar & Reducindo, Mireya & Hernández, Alix & García, Manuel & Gonzalez, Francisco. (2018). An Automated Method for the Evaluation of Breast Cancer Using Infrared Thermography. EXCLI Journal. 17. 989-998. 10.17179/excli2018-1735.
[16] NIRAMAI website. https://www.niramai.com/faqs/
[17] Ojala, Timo, Matti Pietikainen, and Topi Maenpaa. "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns." IEEE Transactions on pattern analysis and machine intelligence 24.7 (2002): 971-987.
[18] Pereira, Eanes Torres, Sidney Pimentel Eleutério, and João Marques Carvalho. "Local binary patterns applied to breast cancer classification in mammographies." Revista de Informática Teórica e Aplicada 21.2 (2014): 32-46.
[19] Raghavendra, U., et al. "An integrated index for breast cancer identification using histogram of oriented gradient and kernel locality preserving projection features extracted from thermograms." Quantitative InfraRed Thermography Journal 13.2 (2016): 195-209.
[20] Silva, L.F.; Saade, D.C.M.; Sequeiros, G.O.; Silva, A.C.; Paiva, A.C.; Bravo, R.S.; Conci A. "A New Database for Breast Research with Infrared Image". Journal of Medical Imaging and Health Informatics 4(1)(2014): 92-100.
[21] Silva 2016. Hybrid analysis for indicating patients with breast cancer using temperature time series. Computer Methods and Programs in Biomedicine tumors using digital infrared imaging. Quantitative InfraRed Thermography Journal. 8. 10.3166/qirt.8.139-148.
[22] Vardasca, R.; Magalhaes, C.; Mendes. "Current State of Machine Learning Classification". J. Biomedical Applications of Infrared Thermal Imaging. Proceedings (2019): 27-46.