10.21611/qirt.2020.078

QIRT PORTO – PORTUGAL 2020, 6-10 JULY

15th Quantitative InfraRed Thermography Conference

Quantitative human interpretation for breast thermography

by R. Gonzalez-Leal*, M. Kurban*, F. J. Gonzalez*, and Orquídea Cruz*

* Eva Tech, Diagonal Patriotismo 12 P-3 Col. Hipódromo Condesa, Ciudad de México 06100, Mexico, raymundo.gonzalez@evacenter.com, kurbanrita@gmail.com, javier.gonzalez@evacenter.com, orquidea.cruz@evacenter.com

Abstract

Studies that evaluate thermography as a method for breast cancer detection have relied on qualitative evaluation, without a well-defined evaluation technique. Some researchers have proposed systems through which interpreters can quantify their findings in thermographic images. The present work reviews and compares human evaluation methods from the literature in terms of their ability to predict cancer status. Also, additional features that can improve the performance of existing methods are identified and tested.

1. Introduction

Currently, there is no standard methodology for the human interpretation of infrared images of the breast. In the Breast Cancer Detection Demonstration Projects (BCDDP), only 5 out of 27 centers had interpreters familiar with infrared thermography. Eighteen months into the project, training was provided, but only 11 centers sent their technicians to the training program. They did not use any scoring system that would allow quantifying results other than indicating whether the thermogram seemed "normal" or "abnormal" [5].

Since then, several studies have proposed methods to evaluate thermograms; one of the most popular ones is the Marseille system, which defines 5 TH categories based on a variety of thermal findings [8]. Papers mention that this method is "specific, objective, and quantitative." This method requires the cold challenge — a one-minute immersion of the hands into cold water before the examination.

The Ville Marie study proposed a quantitative method that consisted of assigning a low numerical value to cases without vascular patterns and with existing but symmetrical patterns or only moderate asymmetry. In their scale, higher values are assigned depending on the number of "abnormal signs" identified, as shown in Table 2. Using this grading scale, the authors found a combined sensitivity of 95% of thermography and mammography, compared to a sensitivity of 85% for mammography alone in 100 cases of DCIS [5].

Table 2. *Ville Marie Infrared (IR) Grading Scale*

Abnormal Signs
1. Significant vascular asymmetry* 2. Vascular anarchy consisting of unusual tortuous or serpiginous vessels that form clusters, loops, abnormal arborization, or aberrant patterns. 3. A 1°C focal increase in temperature (ΔT) when compared to the contralateral site when associated with the area of clinical abnormality. 4. A 2°C focal ΔT versus the contralateral site. 5. A 3°C focal ΔT versus the rest of the ipsilateral breast when not present on the contralateral site. 6. Global breast ΔT of 1.5°C versus the contralateral breast.
Infrared Scale
IR1 = Absence of any vascular pattern to mild vascular symmetry IR2 = Significant but symmetrical vascular pattern to moderate vascular asymmetry, particularly if similar to prior imaging IR3 = One abnormal sign IR4 = Two abnormal signs IR4 = Three abnormal signs
*Unless stable or serial imaging or due to known noncancer causes (e.g., abscess or recent surgery)

Kontos et al. [6] developed another breast thermal imaging scale based on the color gradient between adjacent areas and differences between the two breasts. Images were ranked on a scale from T1 to T5, where T1-2 were indicative of healthy tissue or benign changes, T3 marked lesions with uncertain malignant potential, and T4-5 were cases suspicious or highly suspicious of malignancy. They further defined T1-2 as the absence of focal, non-linear differences in temperature with four or more colors difference from surrounding area and absence of diffuse lesions with six or more colors different from the contralateral breast. Despite being relatively better defined compared to the previously discussed methods, this approach still lacks a clear quantitative definition.

González [2] introduces a thermal score to analyze thermographic findings quantitatively. This score is defined as the sum of the vascularity value and the temperature difference in degrees Celsius at the lesion site compared to the contralateral breast (delta-T). If no known lesion exists, then the highest temperature region in the breast is used when computing the difference in temperature. Gonzalez obtained a significant correlation between thermal score and tumor size [2].

Table 3. *Vascularity Grading Scale*

Score	Explanation
1	Absence of vascular patterns.
2	Symmetrical or moderate vascular patterns.
3	Significant vascular asymmetry.
4	Extended vascular asymmetry in at least one-third of the breast area.

Morales-Cervantes et al. [7] developed an automated method to compute the Gonzalez score [2] and evaluated it in 206 patients, 8 of them with breast cancer. The study found an area under the ROC curve of 0.83 using the thermal score for binary classification of patients. The model predicted all cancer cases to be positive.

As far as the authors are aware, no comparison between quantitative methods to analyze thermal images of the breast has ever been made in the same set of patients. This work compares the methods presented in Gonzalez and Keyserlingk et al. [2,5] because they have a clear quantitative definition and do not require any additional procedures before the examination. We also introduce a new method that accounts for more thermal criteria than the methods mentioned above.

Eventually, machine learning may provide a fully standardized method for the interpretation of breast thermograms. While a machine learning system could be prone to biases due to inappropriate selection of the training data, it would be free from human bias. Hence, at the very least, the predictions would be consistent as long as no changes are made to the model's architecture and its training data and procedure.

Understanding the best strategy that a human can follow in interpreting breast thermograms may provide insights that aid the development of reliable machine learning models. These insights would not necessarily be limited to ideas of manual features for classical machine learning models. They could also inform viable data augmentation policies, model architecture decisions, or help evaluate whether a model is likely to generalize well based on the areas of the image that most influenced its decision.

One of the most critical questions when evaluating a machine learning model for the interpretation of medical images is whether the method exceeds human performance. Answering that question in the case of thermal images in a standardized and replicable manner requires a definition of the appropriate procedure a human should follow in interpreting such images. This work aims at taking the first steps toward using quantitative methods to define such standards.

2. Method

Data comes from the Visual Lab Database for Mastologic Research and the nonprofit organization Eira in Mexico. It consists of 81 samples, 33 of which are confirmed cancer cases. The interpreters did not know anything about the origin of the thermograms.

Images from both centers had the same aspect ratio but different sizes. Visual Lab images have a resolution of 640 x 480 pixels, while Eira images come at 320 x 240 pixels. In order to prevent the interpreters from quickly discerning the two centers apart, Eira images were upscaled using the OpenCV's Lanczos interpolation. We removed patients with visible surgical procedures in their breasts, as interpreters would likely suspect abnormality in the surgically manipulated breast.

The researchers evaluated the thermograms based on the criteria necessary to compute the Gonzalez and the Keyserlingk scores, as well as some other factors that we hypothesized could improve predictions:

1. **Regions of interest:** The number of separate vascularity regions that researchers considered necessary for the analysis. The most common value is 1, with 0 in cases where no asymmetries are present, and values higher than 1 if there are multiple areas of potential abnormality.
2. **Asymmetry size:** Size of the asymmetric vascularity region on a scale from 1 to 10. If there is more than one asymmetrical vascular region, only the larger one is considered.
3. **Asymmetry shape:** How "abnormal" the asymmetry looks to the interpreter on a scale from 1 to 10. Only the "most abnormal" looking vascularity region is considered.
4. **Nipple asymmetry:** A binary value to indicate whether the thermal patterns around the nipples are asymmetrical.
5. **Breast shape asymmetry:** A binary value to indicate whether the breasts significantly differ in shape.
6. **Delta-T:** The temperature difference between the hottest point in the asymmetrical vascularity region and its corresponding location on the opposite breast.
7. **Vascularity:** A vascularity score between 1 and 4. This value combines both asymmetry size and shape. This metric was added to replicate the Gonzalez method [2].
8. **Subjective score:** The interpreter's level of suspicion on a scale from 0 to 10, where 0 represents certainty that the patient is healthy, and 10 indicates certainty that the case is abnormal. The interpreter is asked to give this score based on their previous experience without using any metrics.

3. Results

Two independent researchers scored the images on the metrics presented above. The first researcher, R1, has many years of experience in interpreting medical thermography, while the second researcher, R2, has been freshly trained right before the study. After obtaining manual evaluations, we implemented two quantitative methods [2,5]. It is important to note that our data do not contain any images that satisfy the fifth criteria from the Keyserlingk score [5]. Thus, we omitted this criterion from the final score implementation without risking to lose information. We also constructed an alternative score based on the Gonzalez score. Instead of combining asymmetry shape and size into one vascularity feature, we took them separately since these metrics are more intuitive for humans than the combined score. The delta-T value was added to the combination of these features to get the final score.

To evaluate how similar the evaluations of two researchers are, we calculated the Spearman correlation coefficient. One of the main advantages of the Spearman rank correlation coefficient is that the values can be ordinal, and it does not require approximate normal distributions for the variables.

Table 4. *The table indicates the correlation of evaluations given by two researchers. Spearman coefficient is calculated for each metric on the left. MeanR1 and Mean2 are the mean values of the metrics given by each researcher.*

Feature	Spearman Coefficient	Mean R1	Mean R2
# Regions of Interest	0.26	0.94	1.33
Size of Asymmetry	0.26	2.60	5.02
Vascularity	0.17	1.75	3.01
Shape of Asymmetry	0.21	2.86	6.57
Delta-T	0.67	1.30	1.34
Nipple Asymmetry	0.13	0.48	0.79
Breast Shape Asymmetry	0.18	1.86	5.44
Subjective	0.34	3.85	7.33

Spearman coefficients indicate a moderate-to-low positive correlation for different criteria. Temperature difference (delta-T) has the highest correlation score because this is one of the most straightforward metrics that researchers calculate. Other scores appear to be more interpreter-dependent and, therefore, have a lower correlation. The lowest correlated feature is the Nipple Asymmetry since the nipple is relatively small compared to the rest of the breast, which makes the evaluation harder. However, this metric is essential since an asymmetric nipple can be a sign of a rare form of breast cancer — Paget's disease.

We implemented the scores using researchers' evaluations, normalized the values to the range from 0 to 1, and calculated the Area Under a Curve (AUC) score for each of them. Figure 1 compares the performance of the quantitative scores [2,5] against the subjective score as well as the new score.

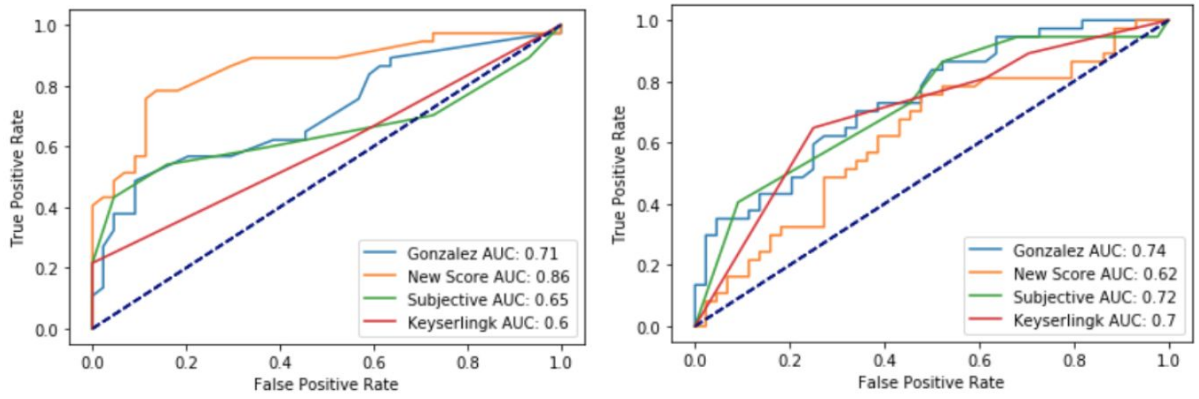

Figure 1. *The AUC scores for different thermal scores as given by the first researcher (left) and the second researcher (right).*

With the AUC score of 0.71 and 0.74, both researchers performed comparably well on the Gonzalez score [2]. The results obtained for the Keyserlingk score were significantly lower. A potential explanation is that while the Keyserlingk score is based on the quantity of the abnormal thermal signs, the Gonzales score also takes into account the magnitude of these abnormalities. Thus, the Gonzalez score provides a more precise understanding of the patient's clinical picture, especially in extreme cases of a significant temperature difference in two breasts that cannot be captured by the Keyserlingk score.

While a more experienced researcher performed exceptionally well on the new score, improving their result by a high margin, leading to the AUC score of 0.86, a less experienced interpreter got the lowest performance of 0.62. To investigate this difference, we carefully examined the evaluations of both researchers to understand their approaches better.

By looking at the mean evaluations given by researchers (Table 4), we can see that the second researcher tends to give much higher values for features used in the score calculation. We explain it by a significant disparity in researchers' experience. While a less experienced researcher had difficulty quantifying certain factors, such as shape and size of asymmetry, the other researcher based evaluations on their prior experience, which led to a more holistic and consistent evaluation. While less correlated features are preferable for machine learning models, we believe that the strategy taken by the first interpreter led to better results since breast cancer manifests itself through several signs that should be considered in combination. For example, a short interview after researchers performed evaluations revealed that a more experienced researcher took delta-T into account when giving scores to the size and shape of asymmetry to get a better understanding of the scale of the change. As a result, the correlation of scores with the BI-RADS status was significantly better compared to that of their colleague.

Additionally, we can see that both the size and shape of asymmetry features were highly correlated with BI-RADS for the first researcher but not for the second. The correlation of vascularity and delta-t to the BI-RADS score is comparable in both cases. It can explain the dramatic discrepancy in the results obtained using the new score compared to the Gonzalez score.

Table 5. *Correlation of metrics to BI-RADS score for R1 and R2.*

Feature	Spearman Coefficient for R1	Spearman Coefficient for R2
Size of Asymmetry	0.50	0.01
Vascularity	0.16	0.09
Shape of Asymmetry	0.62	0.16
Delta-T	0.41	0.49

4. Conclusion

This study makes the first step towards specifying quantitative evaluation standards for infrared thermography. We analyzed two promising scores from the literature introduced in Gonzalez and Keyserlingk et al. [2,5]. The experimental study indicated that the Gonzalez score outperformed the Keyserlingk score as well as entirely subjective

interpretations given by researchers. We also introduced a new score by separating a single vascularity metric into the size and shape of asymmetry measures. This approach leads to a significant increase in performance but only for a researcher with many years of experience. These findings give us a reason to assume that the BCDDP study would probably obtain better results if performed today with experienced interpreters and quantitative methods since subjective scores are not the best and the level of researchers' experience matters. A further study is needed to explore ways of making the new method less interpreter-dependent. For example, a study of the differences in approaches taken by researchers can be used to inform the training of new thermography evaluators.

REFERENCES

[1] Arora, N., Martins, D., Ruggerio, D., Tousimis, E., Swistel, A. J., Osborne, M. P., & Simmons, R. M. (2008). Effectiveness of a noninvasive digital infrared thermal imaging system in the detection of breast cancer. The American Journal of Surgery, 196(4), 523-526.
[2] González, Francisco Javier. "Non-invasive estimation of the metabolic heat production of breast tumors using digital infrared imaging." Quantitative InfraRed Thermography Journal 8.2 (2011): 139-148.
[3] González, Francisco Javier, Raymundo González, and Juan Carlos López. "Thermal contrast of active dynamic thermography versus static thermography." Biomedical Spectroscopy and Imaging 8.1-2 (2019): 41-45.
[4] Kennedy, D. A., Lee, T., & Seely, D. (2009). A comparative review of thermography as a breast cancer screening technique. Integrative cancer therapies, 8(1), 9-16.
[5] Keyserlingk, John R., et al. "Functional infrared imaging of the breast." IEEE Engineering in Medicine and Biology Magazine 19.3 (2000): 30-41.
[6] Kontos, M., Wilson, R., & Fentiman, I. (2011). Digital infrared thermal imaging (DITI) of breast lesions: sensitivity and specificity of detection of primary breast cancers. Clinical radiology, 66(6), 536-539.
[7] Morales-Cervantes, A., Kolosovas-Machuca, E. S., Guevara, E., Reducindo, M. M., Hernández, A. B. B., García, M. R., & González, F. J. (2018). An automated method for the evaluation of breast cancer using infrared thermography. EXCLI journal, 17, 989.
[8] Spitalier, J. M., Kurtz, J. M., Amalric, R., Brandone, H., Ayme, Y., Bressac, C., & Hans, D. (1989). Long-term survival following breast-conserving therapy in comparison to radical surgical treatment: An overview. In Breast Diseases (pp. 276-284). Springer, Berlin, Heidelberg.
[9] Omranipour, Ramesh, et al. "Comparison of the accuracy of thermography and mammography in the detection of breast cancer." Breast Care 11.4 (2016): 260-264.