Reinhard it: Normalization and Classification on HER2 images

Luan Matheus Trindade Dalmazo[1], **Gabriel Silva Hermida**[1], **Sergio Ossamu Ioshii**[2], **Lucas Ferrari de Oliveira**[1]

[1]Departamento de Informática – Universidade Federal do Paraná (UFPR)
Caixa Postal 19.081 – 81.531-980 – Curitiba – PR – Brazil

[2]Pathological Anatomy Laboratory – Erasto Gaertner Hospital

{luantrindade,lferrari,silva.gabriel}@ufpr.br, sergio.ioshii@pucpr.br

Abstract. *HER2-positive breast cancer is one of the most aggressive subtypes and among the most frequently diagnosed. It results from the overexpression of the HER2 protein, which is assessed using the ImmunoHistoChemistry (IHC) score. However, this evaluation is often performed manually, creating opportunities for automation. In this context, this study investigates the impact of the Reinhard technique compared to mean and standard deviation normalization methods to quantify protein levels. The results demonstrate the potential of the proposed approach, achieving an F1-score of 0.89, precision of 0.89, and recall of 0.90, whereas the mean and standard deviation normalization method obtained 0.85 for precision and 0.86 for both F1-score and recall.*

1. Introduction

Breast cancer is highly prevalent in Brazil, with 73,610 new cases estimated by the Brazilian National Cancer Institute (INCA, 2022) for the period ending in 2025. Given this scenario, precise diagnostic tests are essential to guide treatment. Among these, the evaluation of the Human Epidermal Growth Factor Receptor Type 2 (HER2) stands out [Cordeiro 2019], as HER2-positive breast cancer accounts for approximately 20% of all cases. This subtype is characterized by HER2 protein overexpression, which serves both as a biomarker and a prognostic factor. However, its analysis is often performed manually, making it prone to variability and human bias [Yousif et al. 2021].

Hematoxylin and eosin (H&E) staining is the gold standard for visualizing tissue cell structure and morphology. The ImmunoHistoChemistry (IHC) scoring system classifies HER2 expression based on reaction intensity and the number of membrane-positive cells, using a scale from 0 to 3+ [Cordeiro 2019]. In this context, image processing algorithms and machine learning help automate analysis, with studies improving accuracy and early diagnosis [Thakur and Kutty 2019]. However, careful pre-processing is often necessary to enhance model performance by reducing intra-class variability and increasing inter-class distinctions [Wang et al. 2020].

To address classification challenges in HER2-IHC, normalization techniques can help standardize image features. Among them, the Reinhard technique is a promising approach, adjusting the color distribution of the source image to match that of the target image [Reinhard et al. 2001].

This study evaluates the impact of the Reinhard technique on a HER2-IHC dataset using ResNet50 for classification. The assessment will be conducted in comparison with normalization by mean and standard deviation. The main contributions of this work include the proposal of two pipelines, one for each normalization method, and a discussion of the results obtained for each.

2. Related Works

Over the years, various studies have explored methods for classifying HER2 images in breast cancer, often incorporating preprocessing techniques such as Reinhard normalization. One such approach is the modified Reinhard technique proposed by [Panda et al. 2022] for H&E stained histopathology images, which preserves background luminance. This method demonstrated high effectiveness, achieving a Pearson correlation coefficient of 0.96, outperforming other techniques.

Other works have investigated different strategies. The authors [Assolari and de Freitas 2023] focused on Ki-67 quantification, testing three preprocessing methods: no preprocessing, CLAHE, and CLAHE combined with Reinhard normalization. The latter achieved the best results, with a correlation coefficient of 0.83. Additionally, [P et al. 2024] proposed an ensemble deep learning model for HER2 classification using H&E-stained images, achieving an accuracy of 97.84% and an AUC of 100% on the Breast Cancer Immunohistochemistry dataset.

Therefore, it is clear that all the studies highlight the effects of normalization and preprocessing steps on metrics such as correlation and accuracy, emphasizing the impact of stages prior to classification itself. In contrast to the presented studies, this work applies the Reinhard technique for image normalization, comparing it to mean and standard deviation normalization, using the ResNet50 neural network for HER2 image classification.

3. Methodology

The methodology of this work involves splitting the datasets for training and testing, image normalization, and the subsequent training and testing phases. This section provides a detailed explanation of each step and the rationale behind the chosen methods. Figure 1 illustrates the adopted pipeline, showing two analysis flows: one applying the Reinhard technique and the other using the mean and standard deviation, to assess the impact of the method on the evaluated metrics.

3.1. Dataset

The dataset used contains HER2 samples divided into four classes based on the level of HER2 expression. Class 0 consists of 277 images, representing samples with no or very low expression of HER2 (negative). Class 1+ includes 387 images, with low HER2 expression (negative). Class 2+ contains 419 images, which correspond to moderate HER2 expression, while class 3+ comprises 420 images, representing high HER2 expression (positive). The dataset was split in a 3/4 ratio for training and 1/4 for testing, ensuring that images from the same patient were kept in the same set to control potential overfitting.

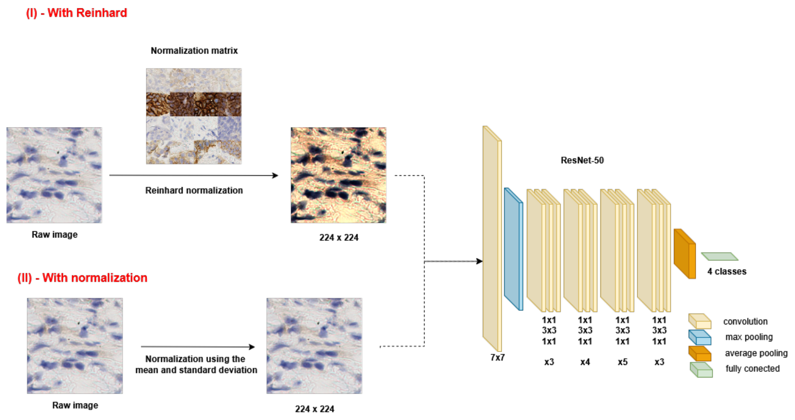

Figure 1. Illustrative diagram of the adopted pipeline.

3.2. Preprocessing

Initially, for the flow that applies the Reinhard technique, a normalization matrix was constructed, as shown in Figure 2, which serves as a 'template', with each row containing images of a specific class. The technique proposed by [Reinhard et al. 2001] adjusts the color characteristics of an image to match those of a target image. This involves converting the image from the RGB color space to LAB (Luminance and Chrominance), using the LMS (Long, Medium, Short) color space as an intermediate step, with the conversion values shown in Table 1. The Reinhard technique is then applied using this matrix as the target, and the normalization process involves calculating the mean and standard deviation of each RGB channel, followed by normalizing the image based on these values. Algorithm 1 presents the flow of instructions for normalization, and the images normalized using the Reinhard technique are shown in Figure 3.

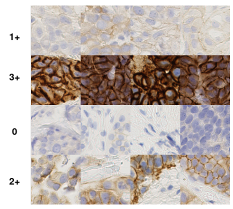

Figure 2. Normalization matrix.

Algorithm 1 Normalization of images using Reinhard transform

Input: List of image paths, target image in LAB color space (normalization matrix)
Output: Normalized images
$l, a, b \leftarrow$ split($target\_lab$)
$target\_mean \leftarrow$ [mean(l), mean(a), mean(b)]
$target\_std \leftarrow$ [std(l), std(a), std(b)]
for each $image\_path$ in $list\_images$ **do**
    $image \leftarrow$ read image from $image\_path$
    $image \leftarrow$ convert($image$, BGR to RGB)
    $normalized\_image \leftarrow$ Reinhard($image, target\_mean, target\_std$)
end for

Table 1. Matrices used for RGB $\rightarrow$ LMS and LMS $\rightarrow$ LAB transformations.

Transformation	Matrix
RGB $\rightarrow$ LMS	$\begin{pmatrix} 0.3811 & 0.5783 & 0.0402 \\ 0.1967 & 0.7244 & 0.0782 \\ 0.0241 & 0.1288 & 0.8444 \end{pmatrix}$
LMS $\rightarrow$ LAB	$\begin{pmatrix} \frac{1}{\sqrt{3}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{6}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -2 \\ 1 & -1 & 0 \end{pmatrix}$

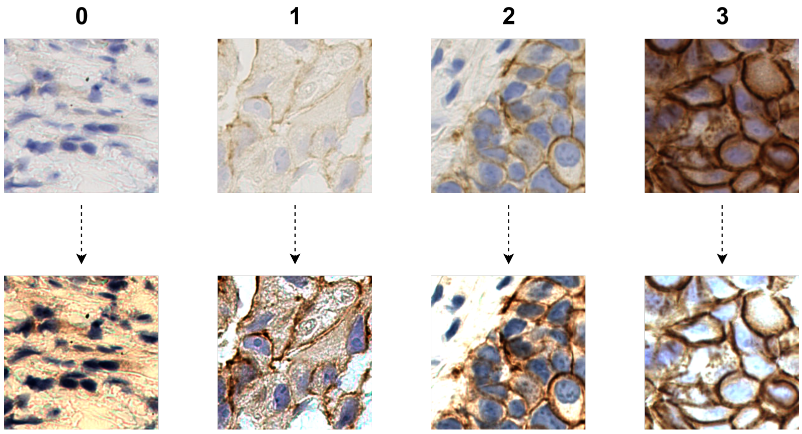

Figure 3. Before and after Reinhard normalization.

In order to facilitate convergence and ensure a similar scale within the data, the flow that does not use the Reinhard technique was normalized, this time using the mean and standard deviation obtained for the image sample, taking into account the number of samples, the number of channels, and the number of pixels, so that the data would be on similar scales, thereby facilitating convergence. For each RGB channel, the mean values were 0.6972 (R), 0.6460 (G), and 0.6218 (B), while the standard deviation values were 0.1117, 0.1153, and 0.1183, respectively.

3.3. Training

For training, the hyperparameters used during the transfer learning of ResNet50 included a batch size of 32, cross-entropy loss as the loss function, and the Adam W optimizer. The learning rate was set to 1×10^{-4}, and the model was trained for 20 epochs. All network weights were frozen, except for the classification layer, which was specifically trained for the dataset in this study. Additionally, the output of the classification layer was modified from 2048 to 4 to match the number of classes in the study. At each epoch, the samples were randomized, and the best model was saved for use in the testing phase. In total, three training sessions were conducted.

3.4. Test

This stage focuses on visualizing the model's performance. The three previously trained models will be evaluated using precision, recall, and F1 score metrics, both as averages and in their macro versions. To assess the consistency of the metrics for each model, the standard deviation will also be calculated.

4. Results and Discussion

The results presented in Table 2 highlight the positive contribution of Reinhard's normalization technique to the classification task, reflected by the improvement in precision, recall, and F1-score metrics. The standard deviation of 0.01, obtained from the three runs, emphasizes the consistency of the achieved results. The improvements in the metrics indicate a reduction in false positives while enhancing the detection of true cases. In the HER2 protein analysis, this improvement is particularly crucial, as it minimizes the risk of incorrect diagnoses, which must be avoided at all costs.

Models performance for each one of the classes is described in Table 3. Despite the aforementioned overall better performance when applying the Reinhard technique to the data, it is shown that it had lower precision when classifying 3+ samples. However, a better recall score means that the Reinhard normalized samples are better classified as true positives, which can be valuable in the context of HER2 classification as it is arguably useful to have a trade-off between precision and recall. Therefore, the lower precision can be compensated by the better recall and F1 score without significant changes.

Table 2. Overall model performance. Bold values indicate improvements observed.

Class	**Precision**	**Recall**	**F1**
Normalized	0.85 ± 0.01	0.86 ± 0.01	0.86 ± 0.01
Reinhard-normalized	**0.89 ± 0.00**	**0.90 ± 0.00**	**0.89 ± 0.00**

Table 3. Measurements of the macrometrics on normalized and Reinhard-normalized data. Bold values indicate improvements observed.

Class	**Reinhard-normalized**			**Normalized**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
0	**0.84 ± 0.01**	**0.98 ± 0.01**	**0.90 ± 0.01**	0.81 ± 0.01	0.96 ± 0.02	0.88 ± 0.01
1+	**0.88 ± 0.01**	**0.81 ± 0.02**	**0.84 ± 0.01**	0.82 ± 0.01	0.73 ± 0.04	0.77 ± 0.01
2+	**0.89 ± 0.01**	**0.85 ± 0.02**	**0.87 ± 0.01**	0.81 ± 0.02	0.82 ± 0.02	0.81 ± 0.00
3+	0.93 ± 0.01	**0.96 ± 0.00**	0.95 ± 0.00	**0.97 ± 0.00**	0.95 ± 0.00	**0.96 ± 0.00**

Figure 4 presents the confusion matrices for the two tested scenarios, aiming to further illustrate the classification performance. As previously mentioned, the model using the Reinhard technique is more successful in correctly classifying true positives, especially for the 3+ class, although it shows minimal confusion between this class and the 2+ class. This confusion may be due to similarities between some images from these two classes. Additionally, for the intermediate expression levels of 1+ and 2+, both models exhibit significant errors, with confusion with other classes. However, the model using the Reinhard normalization technique still demonstrates better performance.

5. Conclusion

In conclusion, Reinhard normalization significantly enhances performance metrics compared to method without normalization. This image preprocessing technique is effective due to its low computational cost and the improved results it yields. Future work will

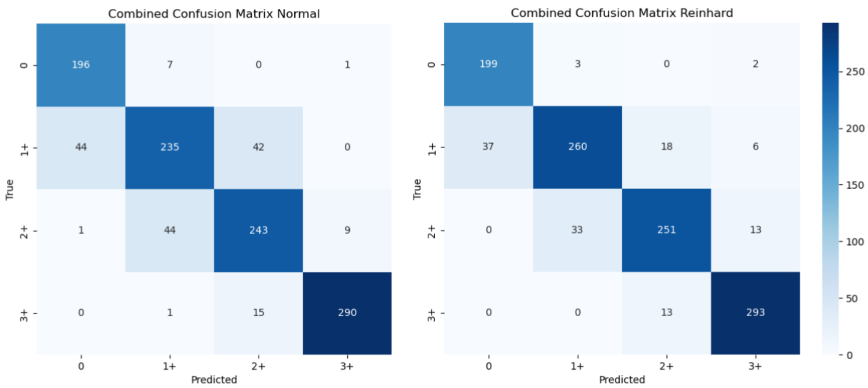

Figure 4. Confusion matrices for both tested scenarios .

focus on evaluating its application to different datasets and neural network models to further validate the findings. Additionally, the proposed pipeline is expected to contribute to disease diagnosis by reducing testing time and offering a precise, automated method for IHC. The codes used in this study are available at this GitHub Repository.

References

Assolari, C. L. and de Freitas, P. M. (2023). Automação do método de quantificação do Índice proliferativo ki-67 do câncer de mama.

Cordeiro, C. Q. (2019). An automatic patch-based approach for her-2 scoring in immunohistochemical breast cancer images. Master's thesis, Universidade Federal do Paraná, Curitiba, PR.

P, P. G., Senapati, K., and Pandey, A. K. (2024). A novel decision level class-wise ensemble method in deep learning for automatic multi-class classification of her2 breast cancer hematoxylin-eosin images. *IEEE Access*, 12:46093–46103.

Panda, S., Jangid, M., and Jain, A. (2022). Enhancing background luminance for colorectal cancer h and e stained images using modified reinhard technique. In *2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, pages 129–133.

Reinhard, E., Adhikhmin, M., Gooch, B., and Shirley, P. (2001). Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41.

Thakur, V. and Kutty, R. V. (2019). Recent advances in nanotheranostics for triple negative breast cancer treatment. *Journal of Experimental Clinical Cancer Research*, 38(1):430.

Wang, Y., Lei, B., Elazab, A., Tan, E.-L., Wang, W., Huang, F., Gong, X., and Wang, T. (2020). Breast cancer image classification via multi-network features and dual-network orthogonal low-rank learning. *IEEE Access*, PP:1–1.

Yousif, M., Huang, Y., Sciallis, A., Kleer, C. G., Pang, J., Smola, B., Naik, K., McClintock, D. S., Zhao, L., Kunju, L. P., Balis, U. G. J., and Pantanowitz, L. (2021). Quantitative image analysis as an adjunct to manual scoring of er, pgr, and her2 in invasive breast carcinoma. *American Journal of Clinical Pathology*, 157(6):899–907.