Automatic ER and PR scoring in Immunohistochemistry H-DAB Breast Cancer images

Johanna Elisabeth Rogalsky[1], **Sergio Ossamu Ioshii**[2], **Lucas Ferrari de Oliveira**[1]

[1] Universidade Federal do Paraná (UFPR)
Departamento de Informática
Curitiba, PR – Brazil

[2]Pontifícia Universidade Católica do Paraná (PUCPR)
Pós-Graduação de Tecnologia em Saúde da PUCPR
Curitiba, PR – Brazil

johanna.elisabeth8@gmail.com

sergio.ioshii@pucpr.br,lferrari@inf.ufpr.br

Abstract. *Breast Cancer (BC) is the most frequently diagnosed cancer for women. This way, the Brazilian Unified Health System (SUS) focuses on studying the disease and improving all the steps involved in dealing with BC. The presence or absence of the Estrogen Receptor (ER) and the Progesterone Receptor (PR), which define invasive subtypes, is detected through Immunohistochemistry (IHC). One way to assist the manual assessment of pathologists and histopathologists is to develop automatic scoring systems. Fortunately, digital pathology is increasingly achieving higher agreement with the pathologist. Therefore we create an automatic scoring system composed of image preprocessing, feature extracting, and classification achieves a 69% f-score rate.*

1. Introduction

Cancer is a group of diseases characterized by the uncontrolled growth and spread of abnormal cells that can lead to death. Unfortunately, only the increasing risk factors are known, while the associated root causes remain uncharted [Society 2018]. This way, classifying cancer as a leading cause of death and an obstacle to the increase in life expectancy worldwide [Sung et al. 2021].

In 2020, the GLOBOCAN 2020 estimated Breast Cancer (BC) as the most incident cancer with 2.3 million new cases and 685 thousand deaths, surpassing lung cancer when compared with the GLOBOCAN 2018. Furthermore, BC was the most frequent cancer diagnosed for women in almost every country and the most frequent cause of death in a little more than half of them [Sung et al. 2021].

In Brazil, BC was the cancer type with the highest mortality rate. In addition, the GLOBOCAN 2020 Brazilian numbers estimate were 88,492 new cases and 20,725 deaths, besides the 299,542 cases for 5-year BC prevalence [Ferlay et al. 2020]. Resulting in it being the cancer type with the highest rates with a magnitude two to three times greater than the second most frequent [Instituto Nacional de Câncer 2019b].

To have a better knowledge of the disease course and its molecular characteristics, the Brazilian Unified Health System (SUS) has been improving the diagnosis and the local and systemic treatments for BC. Its *Linha de Cuidado no Câncer de Mama* has, since the year 2000, been the source of information and research material. That consists of strategic planning that organizes actions and precautions for prevention, early detection, diagnostic investigation, and palliative care.

Because there are various histological and molecular invasive subtypes, a molecular pattern evaluation is important to acknowledge both prognosis and response to specific treatments or therapies. That, consists of qualifying the Estrogen Receptor (ER) and the Progesterone Receptor (PR) biomarkers through Immunohistochemistry (IHC) tests [Instituto Nacional de Câncer 2019a] using the Positive cells Intensity score (IS), with 0: no staining, 1+: weak positive staining, 2+: moderate positive staining and 3+: strong positive staining, as possible values.

Traditionally, this quantification is still done manually [Liu et al. 2016]. This way, it depends on the pathologist's or histopathologists experience and professional background, being an expensive, tedious, and time-consuming task that can over-fatigue the professional, causing errors and misdiagnosis [Han et al. 2017]. Also suffering from intra- and interobserver variability [Robertson et al. 2018]

Fortunately, digital pathology scanning systems and image analysis tools have become more popular in recent years [Tollemar et al. 2018]. Therefore, considering the increasing need to develop automated imaging systems to support experts as a second opinion in the prognosis and diagnosis and the necessity for correct and personalized treatment, we propose investigating an automated method for BC ER/PR scoring in Whole Slide Image (WSI).

2. Related Work

QuPath [Bankhead et al. 2017, Bankhead et al. 2018] is a free and publicly bioimage analysis software designed to be a user-friendly and open-source digital pathology and whole slide image analysis solution for desktop. The authors applied their system for T-cell markers CD3 and CD8 analyzing in [Bankhead et al. 2017]. In [Bankhead et al. 2018], the authors applied their system to score five BC biomarkers: ER, PR, HER2, Ki-67 and p53. Utilizing stain estimation, Tissue MicroArray (TMA) dearraying, cell segmentation, feature computation, and tumor cell identification, each core in the images automatically received a score. When comparing with the pathologists' score, the system gives a mean 0.94 AUC score, indicating high agreement and correlation.

ImmunoRatio [Tuominen et al. 2010] is a quantitative image analysis software of the ER, PR and Ki-67 biomarkers, available as an ImageJ plugin. It takes an immunostained cellular image and performs blankfield correction or background subtraction. Then this image passes through color deconvolution to separate the stains by color. Both component images become pseudo-colored result images when processed with filtering, adaptive thresholding, nucleus segmentation, and small particles discard and overlayed onto the source image. That way, the Diaminobenzidine (DAB)/total nuclear area percentage can be pixel counted. For validation, the creators tested their software by applying it on the three biomarkers achieving an r = 0.97 and r = 0.98 correlation with visual cell nuclei counting, when with and without the camera adjustment wizard, respectively.

In the [Mouelhi et al. 2018] work, the authors developed a software able to segment and classify cancer nuclei in IHC images in order to provide quantitative evaluation of ER or PR status. The workflow had two stages: cell nuclei segmentation and cancer nuclei classification using histogram equalization for contrast enhancement and background elimination and adaptive morphological criterion to highlight the cancerous nuclei. Also, four maximal color separation techniques compute the intensity images, allowing to measure the intensity score and calculate the percentage of positive staining for subsequent Allred scoring. The evaluation achieved a 98% rate for sensitivity and accuracy, for both detected nuclei and image cancer scoring over the truths provided by experienced pathologists, showing the best correlation with the expert's score (Pearson's correlation coefficient = 0.993, p-value $<$0.005).

The authors of the [Paulik et al. 2017] work developed and validated an image analysis application aiming it to be a robust nuclear biomarker detector in WSIs. The first step was to extract the image's nuclear channel from the background information using its average intensity. After this, an adaptive thresholding method homogenized the background intensity for a maxima finding algorithm to detect the center cell points and separate them. Each point receives a unique ID for the watershed algorithm to do the segmentation. Also, a hole-filling method is applied utilizing nuclear segmentation and background objects labeling and filtering. The correlation between manual and algorithmic nuclear segmentation reached a good result, with r = 0.99323, n = 30, and P $<$0.05, and precision and recall rates achieved 90.23 ± 4.29% and 88.23 ± 4.84%, respectively.

3. Proposed approach

3.1. Dataset

The dataset created in this work is composed of BC IHC slides from 135 patients from the local (Curitiba, Paraná, Brazil) hospital, whose Ethics Committee on Research approved this work. These patients were tested for both ER and PR, totalizing 270 digitized WSI images made using objective lens of 40x magnification. Also, all images were divided into patches with 400x300 pixels to facilitate the manipulation by the proposed algorithms and had the diagnosis withdrawn from the patient records having the IS score.

3.2. Pre-processing

Based on the work of [Mouelhi et al. 2018], we pre-processed our images with contrast enhancement and thresholding methods, both using local and adaptive approaches. For contrast enhancement, a histogram equalization limits the contrast amplification, reducing noise amplification. And in this case, the method uses the luminance difference between objects and background to enhance.

For thresholding, the first step is to smooth the image at a color level, mainly to erode some of the small color areas, and then overlay the resulting mask with the enhanced image. Then, in order to separate objects by color and remove any remaining background, the image in the HSV color space passes through color separation using the H (hue) channel as a guide, once it is the channel that defines the pure color.

3.3. Feature Extraction

The feature extraction step is expected to describe the images and use them to represent the classes in a discriminate way. That is, make the images enable the classes to be recognizable and distinguishable. To do so, we create one, two, and three-dimensional histograms by using one, two, and three image channels, respectively, in the HSV color space.

It is worth mentioning that histograms account the channel's intensity frequency. So a one-dimensional histogram can be defined as in $h(r_k) = n_k$, where r_k denotes the intensity/gray level with *k* going from 0 to $L-1$, n_k represents the number of pixels in the image *f(x, y)* with intensity r_k and *bins* subdivide the intensity scale [Gonzalez and Woods 2018].

This way, we can define a two-dimensional histogram as $h(r_k, s_l) = n_{k,l}$, where r_k and s_l correspond to the intensities from the two channels with *k* and *l* going from 0 to L_1 and L_2, respectively. And n_k represents the number of pixels in the image $f(x, y)$ that have r_k and s_l intensities as indexes of the matrix in which the histogram is stored. Following the same logic, a three-dimensional histogram can also be created.

3.4. Classification

In order to classify the images, we pass the feature vectors (histograms) sets to the Support Vector Machine (SVM) classifier after normalizing them into a [-1, 1] interval. And to find its best hyper-parameters for the training, we applied the grid search method on top of the dataset evaluated at the time (ER examples and PR examples).

The strategy used for training and testing is the k-fold data separation named *leave-one-patient-out*, meaning that the classifier is trained *N* times with *N - 1* patients and tested with 1, where *N* is the number of patients. And the decision function used is the *one-vs-rest*, that creates *n* classifiers, where *n* is the number of classes.

SVM is a machine learning classifier that separates the training set received as feature vectors by mapping it into a high dimensional feature space and constructing a linear decision surface that ensures the generalization inside it. Consequently, it searches for an optimal hyperplane that separates and generalizes the data well.

So, computationally, the support vectors, which are a few samples taken out of the training set, determine this margin. From this, it is possible to see that SVM is a binary classifier, although a multi-class problem can use it by mapping the training data into an N-dimensional space [Cortes and Vapnik 1995].

4. Results and Discussion

Purposing to evaluate the proposed approach and its experiments, we use the f-score metric extracted from the confusion matrix. As the four IS classes can also be described as negative, borderline, and positive, corresponding to 0/1+, 2+, and 3+, respectively, we did the same experiments for both four and three classes divisions using the images, as mentioned before, in the HSV color space. Table 1 exposes both class division distributions.

The experiments consisted of constructing seven base histograms (Table 2) and use them alone or concatenated (respecting the dimension) as input for the SVM classifier. This way, we had a total of 15 histograms combinations: seven made of the one-dimensional histograms (H, S, V, H_S, S_V, V_H, H_S_V), seven made of the two-dimensional histograms (HS, SV, VH, HS_SV, SV_VH, VH_HS, HS_SV_VH), and the last three-dimensional histogram (HSV).

Table 1. IS classes distribution in the dataset with four (1a) and three (1b) classes.

(a)

IS	Rec ER	Rec PR
0	34	50
1+	13	4
2+	14	33
3+	74	48

(b)

IS	Rec ER	Rec PR
0/1+	47	54
2+	14	33
3+	74	48

Table 2. Base histograms.

Dimensions	Histogram	Bins
1	H	30
	S	32
	V	32
2	HS	30*32
	SV	32*32
	VH	32*32
3	HSV	30*32*32

Using these histograms we performed the experiments and discuss the results for the ER and PR biomarkers IS score separately, since they are two different markers. In addition, we extracted three different sets of histograms: the first one (A) using only the color separation technique; the second one (B) using histogram enhancement, thresholding and color separation; and the third one (C) using the same methods as the former one ((B)), but to extract the histogram only of the positive stained cells, which are the ones with a brown coloration. Figure 1 shows one example for each set.

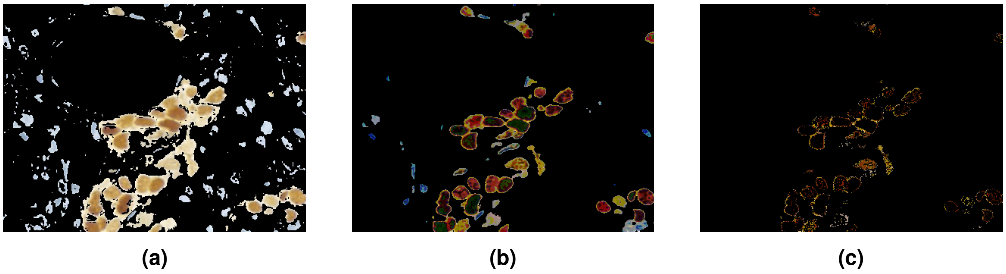
(a) (b) (c)

Figure 1. 1a Image example from the (A) histogram set. 1b Image example from the (b) histogram set. 1c Image example from the (C) histogram set

4.1. ER Image Set

Table 3 shows the f-score for all of the 15 histogram combinations regarding the three histogram sets, A, B, and C, mentioned before, divided into four (4 clss) and three (3clss) classes. Firstly, set B of histograms notably is the one with a slightly higher average rate, which indicates that it could be the best approach for these images, despite the best f-score belonging to set A.

Table 3. F-score from A, B, and C sets of histograms from the ER biomarker.

	A		B		C	
Histogram	4 clss	3clss	4 clss	3clss	4 clss	3clss
H	0.5150	0.6067	0.4953	0.6743	0.4802	0.6471
S	0.4196	0.6351	0.4014	0.6689	0.4813	0.6934
V	0.4196	0.6351	0.4828	0.6299	0.4187	0.6241
H_S	0.5259	0.6443	0.5191	0.6126	0.5384	0.6522
S_V	0.4240	0.6180	0.4861	0.6524	0.4418	0.6424
V_H	0.5259	0.6443	0.5029	0.6414	0.4402	0.6204
H_S_V	0.4906	**0.6986**	0.5160	0.6430	0.4823	0.6466
HS	0.5061	0.6405	**0.5559**	0.6857	**0.5573**	**0.6951**
SV	0.5184	0.6945	0.5147	0.6560	0.4502	0.6601
VH	0.5363	0.6357	0.5060	0.6346	0.5132	0.6341
HS_SV	0.5574	0.6934	0.5440	0.6579	0.4577	0.6333
SV_VH	0.4946	0.6604	0.5082	0.6634	0.4503	0.6323
VH_HS	0.5051	0.6506	0.5579	0.6615	0.5230	0.6053
HS_SV_VH	**0.5876**	0.6826	0.5376	0.6798	0.4531	0.6587
HSV	0.5049	0.6710	0.5445	**0.6937**	0.4133	0.5717
Average	0.5020	0.6540	0.5082	0.6557	0.4778	0.6460

In sets A, B, and C, the best results come from when using the three classes division because the samples are better distributed. When looking at the confusion matrices in Tables 4a, 5a, and 6a, the classes 0 and 3+ are the ones less confused with the others. Which can also be seen in Tables 4b, 5b, and 6b, confirming that the classes with more examples are the ones that the classifier learns the best. Moreover, precisely for having fewer samples, the borderline class (2+) is the worst learn.

Table 4. Average confusion matrix relative to all ER histogram combinations using (4a) four and (4b) three classes in set A.

(a)

Actual class	Predicted			
	0	1+	2+	3+
0	21	5	3	5
1+	5	5	2	1
2+	3	2	4	5
3+	8	6	4	56

(b)

Actual class	Predicted		
	0/1+	2+	3+
0/1+	38	3	5
2+	4	5	4
3+	14	4	57

4.2. PR Image Set

Table 7 presents the metric from each one of the 15 combinations in each set (A, B, and C) when using the four original (4 clss) and the three (3 clss) classes. It is verifiable that set C is the one with the two best f-scores and with the higher average f-score rates, meaning

Table 5. Average confusion matrix relative to all ER histogram combinations using (5a) four and (5b) three classes in set B.

(a)

Actual	Predicted			
class	0	1+	2+	3+
0	18	9	2	4
1+	4	6	2	2
2+	2	2	5	5
3+	5	5	5	58

(b)

Actual	Predicted		
class	0/1+	2+	3+
0/1+	39	3	5
2+	4	5	5
3+	11	5	59

Table 6. Average confusion matrix relative to all ER histogram combinations using (6a) four and (6b) three classes in set C.

(a)

Actual	Predicted			
class	0	1+	2+	3+
0	20	3	2	9
1+	4	3	0	6
2+	2	0	2	10
3+	4	4	3	63

(b)

Actual	Predicted		
class	0/1+	2+	3+
0/1+	37	3	6
2+	4	5	5
3+	8	6	60

this could be the best approach for these images. Furthermore, the three class division enables the classifier to better learning of the classes.

Table 7. F-score from A, B and C sets of histograms from the PR biomarker.

	A		B		C	
Histogram	4 clss	3clss	4 clss	3clss	4 clss	3clss
H	0.4400	0.5785	0.4130	**0.6327**	0.4661	0.5665
S	0.3974	0.5642	0.3909	0.5189	0.4247	0.5506
V	0.3974	0.5642	0.3608	0.5059	0.4070	0.6071
H_S	0.4064	0.5304	0.4017	0.5217	0.4534	0.5547
S_V	0.4015	0.5114	0.3624	0.5262	0.4437	0.5442
V_H	0.4064	0.5304	0.3415	0.5217	0.4312	0.6035
H_S_V	0.4158	0.5304	0.4069	0.5251	0.4531	0.6211
HS	0.4436	0.5340	0.4516	0.5588	**0.4703**	**0.6646**
SV	0.4665	0.5446	0.3815	0.5751	0.4014	0.5792
VH	**0.4686**	0.5540	0.3938	0.5656	0.4479	0.5385
HS_SV	0.4685	0.5620	0.4240	0.5783	0.4099	0.5783
SV_VH	0.4589	0.5653	0.3796	0.5783	0.3858	0.5783
VH_HS	0.4402	0.5653	**0.4468**	0.5720	0.4519	0.5578
HS_SV_VH	0.4382	0.5718	0.4231	0.5783	0.4291	0.5783
HSV	0.3979	**0.5814**	0.4118	0.5814	0.4144	0.5783

Table 7 continued from previous page

Average	0.4326	0.5523	0.3993	0.5594	0.4361	0.5792

When observing the confusion matrices in Tables 8a, 9a, and 10a, the four class division simply does not enable the classifier to learn the 1+ class. But when looking in Tables 8b, 9b, and 10b, the joining with class 0 brings its learning up. Unfortunately in comparison, the borderline class (2+) goes from having almost half samples correctly classified to almost none correct predictions. This happens, because it is the class with the fewer amount of examples, which is the same that happens to class 1+ in the four class division.

Table 8. Average confusion matrix relative to all PR histogram combinations using (8a) four and (8b) three classes in set A.

(a)

Actual	Predicted			
class	0	1+	2+	3+
0	36	4	6	5
1+	2	0	1	1
2+	9	2	16	7
3+	11	1	13	24

(b)

Actual	Predicted		
class	0/1+	2+	3+
0/1+	48	5	1
2+	3	3	27
3+	3	7	39

Table 9. Average confusion matrix relative to all PR histogram combinations using (9a) four and (9b) three classes in set B.

(a)

Actual	Predicted			
class	0	1+	2+	3+
0	35	5	7	4
1+	3	0	1	0
2+	12	2	12	8
3+	10	2	11	25

(b)

Actual	Predicted		
class	0/1+	2+	3+
0/1+	50	3	1
2+	2	2	30
3+	2	6	41

Table 10. Average confusion matrix relative to all PR histogram combinations using (10a) four and (10b) three classes in set C.

(a)

Actual	Predicted			
class	0	1+	2+	3+
0	38	3	5	4
1+	3	0	1	0
2+	9	2	11	11
3+	7	1	10	29

(b)

Actual	Predicted		
class	0/1+	2+	3+
0/1+	48	4	2
2+	4	5	24
3+	3	6	39

5. Conclusion

Finally, BC is a big threat to women's health and, preferably, has to be detected and treated in its early stages. For this, the pathologists manually assess the slides in a time-consuming and tedious work, quantifying and qualifying the slides that contain stained biomarkers, such as ER and PR.

From the results discussed in the previous section, the proposed approach fails to achieve an f-score rate higher to 70%, indicating that the SVM classifier did not learn all classes equally. Considering the distribution of the samples in the classes and analyzing the confusion matrices, a more even distribution of samples could bring the rate up.

To try to resolve this problem, we propose developing a cell segmentation, counting, and classification method as future work. Together with it, we intend to have the diagnosis withdrawn from the patient records revised by a specialist (pathologist/histopathologist).

References

Bankhead, P., Fernández, J. A., McArt, D. G., Boyle, D. P., Li, G., Loughrey, M. B., Irwin, G. W., Harkin, D. P., James, J. A., McQuaid, S., et al. (2018). Integrated tumor identification and automated scoring minimizes pathologist involvement and provides new insights to key biomarkers in breast cancer. *Laboratory Investigation*, 98(1):15.

Bankhead, P., Loughrey, M. B., Fernández, J. A., Dombrowski, Y., McArt, D. G., Dunne, P. D., McQuaid, S., Gray, R. T., Murray, L. J., Coleman, H. G., et al. (2017). Qupath: Open source software for digital pathology image analysis. *Scientific reports*, 7(1):16878.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Ferlay, J., Ervik, M., Lam, F., Colombet, M., Mery, L., Piñeros, M., Znaor, A., Soerjomataram, I., and Bray, F. (2020). Global cancer observatory: Cancer today. Lyon, France: International Agency for Research on Cancer. Available from: https://gco.iarc.fr/today and https://gco.iarc.fr/today/data/factsheets/populations/76-brazil-fact-sheets.pdf, accessed in 05 March 2020.

Gonzalez, R. and Woods, R. (2018). *Digital Image Processing*. Pearson.

Han, Z., Wei, B., Zheng, Y., Yin, Y., Li, K., and Li, S. (2017). Breast cancer multi-classification from histopathological images with structured deep learning model. *Scientific reports*, 7(1):4172.

Instituto Nacional de Câncer (2019a). Breast cancer in brazil: synthesis of information. https://www.inca.gov.br/sites/ufu.sti.inca.local/files//media/document//a_situacao_ca_mama_brasil_2019.pdf. Accessed: 2019-10-23.

Instituto Nacional de Câncer (2019b). Estimate/2020 – cancer incidence in brazil. https://www.inca.gov.br/sites/ufu.sti.inca.local/files//media/document/

/estimativa-2020-incidencia-de-cancer-no-brasil.pdf. Accessed: 2021-03-08.

Liu, J., Qiu, G., and Shen, L. (2016). Luminance adaptive biomarker detection in digital pathology images. *Procedia Computer Science*, 90:113–118.

Mouelhi, A., Rmili, H., Ali, J. B., Sayadi, M., Doghri, R., and Mrad, K. (2018). Fast unsupervised nuclear segmentation and classification scheme for automatic allred cancer scoring in immunohistochemical breast tissue images. *Computer methods and programs in biomedicine*, 165:37–51.

Paulik, R., Micsik, T., Kiszler, G., Kaszál, P., Székely, J., Paulik, N., Várhalmi, E., Prémusz, V., Krenács, T., and Molnár, B. (2017). An optimized image analysis algorithm for detecting nuclear signals in digital whole slides for histopathology. *Cytometry Part A*, 91(6):595–608.

Robertson, S., Azizpour, H., Smith, K., and Hartman, J. (2018). Digital image analysis in breast pathology—from image processing techniques to artificial intelligence. *Translational Research*, 194:19–35.

Society, A. C. (2018). Global cancer facts & figures 4th edition.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249.

Tollemar, V., Tudzarovski, N., Boberg, E., Törnqvist Andrén, A., Al-Adili, A., Le Blanc, K., Garming Legert, K., Bottai, M., Warfvinge, G., and Sugars, R. (2018). Quantitative chromogenic immunohistochemical image analysis in cellprofiler software. *Cytometry Part A*, 93(10):1051–1059.

Tuominen, V. J., Ruotoistenmäki, S., Viitanen, A., Jumppanen, M., and Isola, J. (2010). Immunoratio: a publicly available web application for quantitative image analysis of estrogen receptor (er), progesterone receptor (pr), and ki-67. *Breast cancer research*, 12(4):R56.