IHC Classification in Breast Cancer H&E Slides with a Weakly-Supervised Approach

Sara P. Oliveira[1,2] (sara.i.oliveira@inesctec.pt)
João Ribeiro Pinto[1,2] (joao.t.pinto@inesctec.pt)
Tiago Gonçalves[1,2] (tiago.f.goncalves@inesctec.pt)
Hélder P. Oliveira[1,3] (helder.f.oliveira@inesctec.pt)
Jaime S. Cardoso[2,1] (jaime.cardoso@inesctec.pt)

[1] INESC TEC, Porto, Portugal
[2] FEUP, Porto, Portugal
[3] FCUP, Porto, Portugal

Abstract

Human epidermal growth factor receptor 2 (HER2) evaluation commonly requires immunohistochemistry tests on breast cancer tissue, in addition to the standard haematoxylin and eosin (H&E) staining. Additional costs and time spent on further testing might be avoided if HER2 overexpression could be inferred from H&E slides, as a preliminary indication of the IHC result. We propose a framework that separately processes H&E slide tiles and outputs an IHC label for the whole slide. The network was trained on slides from the HER2 Scoring Contest dataset (HER2SC) and tested on two disjoint subsets of slides from the HER2SC database and the TCGA-TCIA-BRCA (BRCA) collection. The proposed method attained 83.3% classification accuracy on the HER2SC test set and 53.8% on the BRCA test set. Although further efforts should be devoted to achieving improved performance, the obtained results suggest that it is possible to perform HER2 overexpression classification on H&E tissue slides.

1 Introduction

Breast cancer (BCa) is the most commonly diagnosed cancer and the leading cause of cancer-related deaths among women worldwide. However, over the most recent years, despite the increasing incidence trends, the mortality rate has significantly decreased. Among other factors, this results from better treatment strategies that can be delineated from the assessment of histopathological characteristics [1, 2].

The analysis of tissue sections of cancer specimens obtained by biopsy commonly starts with haematoxylin and eosin (H&E) staining, which is usually followed by immunohistochemistry (IHC), a more advanced staining technique used to highlight specific protein receptors, such as the HER2 [3]. In fact, the overexpression of HER2 is observed in 10%–20%[4] of BCa cases and has been associated with aggressive clinical behaviour and poor prognosis [5]. However, these cases have a better response to targeted therapies and consequent improvements in healing and overall survival [5].

The current guidelines [6], revised by the American Society of Clinical Oncology/College of American Pathologists (ASCO/CAP), in 2018, indicate the following scoring criteria for HER2 IHC:

- IHC 0+: no staining or incomplete, faint/barely perceptible membrane staining in 10% of tumour cells or less;
- IHC 1+: incomplete, faint/barely perceptible membrane staining in more than 10% of tumour cells;
- IHC 2+: weak to moderate complete membrane staining in more than 10% of tumour cells;
- IHC 3+: circumferential, complete, intense membrane staining in more than 10% of tumour cells.

Moreover, cases scoring 0+ or 1+ are classified as HER2 negative, while cases with a score of 3+ are classified as HER2 positive. Cases with score 2+ are classified as equivocal and are further assessed by *in situ* hybridization (ISH), to test for gene amplification [6].

Despite the efficiency of IHC and ISH, the additional cost and time spent on these tests might be avoided if all the information needed to infer the HER2 status could be extracted only from H&E slide, as a preliminary indication of the IHC result. However, to the extent of our knowledge, the task of predicting HER2 status on H&E slides has not yet been addressed in the literature, except for a recent challenge[1].

[1] ECDP2020 HEROHE Challenge: `https://ecdp2020.grand-challenge.org`

2 Methodology

The proposed method (Fig. 1) comprises a CNN, pre-trained for the task of HER2 scoring of IHC tiles. The pre-trained parameters are then transferred to the task of HER2 status prediction on H&E tiles, to provide the network with some knowledge of the tissue structures' appearance. Individual tile scores are then combined in a single label for the whole slide.

2.1 Data Preprocessing

For the IHC slides of classes 2+ and 3+, the preprocessing begins with automatic tissue segmentation with Otsu's thresholding obtaining the regions with more intense staining, that correspond to the HER overexpression areas. For slides of classes 0+ and 1+, the segmentation consists of simple removal of pixels with the greatest HSV value intensity, corresponding to background pixels, which do not contain essential information to the problem. These processes, which are performed at $32\times$ downsampled slides, return the masks used in tile extraction. Tiles with size 256×256 are extracted from the slide with original dimensions (without downsampling), provided they are completely within the mask region. These tiles are converted from RGB to HSL colour space, of which only the lightness channel is used. Each tile inherits the class from the respective slide (examples in Fig. 2a–d), turning the learning task into a weakly-supervised problem.

According to the ASCO/CAP guidelines for IHC evaluation, the diagnosis is performed based only on the tumoral region of the slides. Hence, the preprocessing of H&E slides begins with an automatic invasive tissue segmentation with the HASHI method [10, 11]. The segmentation mask is then used to generate H&E tiles (example in Fig. 2e), extracted and processed according to the abovementioned methodology.

2.2 IHC Tile Scoring & H&E Slide Classification

The CNN architecture for the IHC tile scoring consists of 4 convolutional layers (16, 32, 64 and 128 filters, respectively, with ReLU activation). The first layer has a 5×5 kernel, while the remaining have 3×3 kernels. Each convolutional layer is followed by a pooling layer (a max-pooling function without overlap, with kernel 2×2). The network is topped with three fully-connected layers, with 1024, 256, and 4 units, respectively. The first two have ReLU activation, while the third is followed by softmax activation for the output of probabilities for each class.

The network parameters pre-trained with IHC tiles were used as initialization for HER2 status classification on H&E tiles. To achieve a single prediction per tile instead of four, as it was initially trained for on the IHC setting, a soft-argmax activation [12] replaces the softmax activation.

The output scores are then sorted from 3+ to 0+ and the tiles corresponding to the 15% highest ones are selected for the aggregation process. This percentage was chosen to limit the information given to the aggregation network, while still including and barely exceeding the reference 10% of tumour area considered in the HER2 scoring guidelines.

The score aggregation is performed by a multilayer perceptron (MLP), composed of 4 layers, with 256, 128, 64, and 2 neurons, respectively. All layers are followed by ReLU activation and the last one is followed by softmax. Since the input dimension M of the MLP is fixed (we set $M = 300$ to limit memory cost), for images where 15% of the number of tiles exceeds M, they are downsampled to M using evenly distributed tile selection. In cases where 15% of the number of tiles is lower than M, tiles are extracted with overlap, to guarantee that M tiles can be selected. The MLP will process these M HER2 scores and output a single HER2 status label for the respective slide.

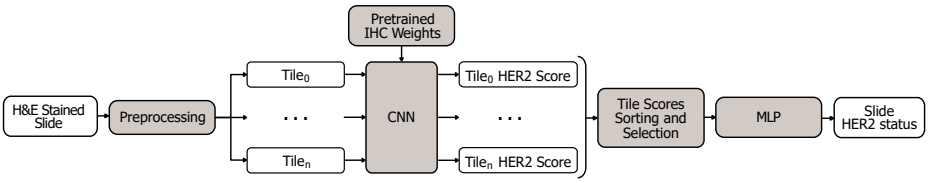
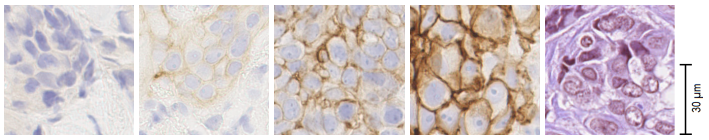


Figure 1: The proposed approach for weakly-supervised HER2 status classification on BCa H&E slides.



(**a**) (**b**) (**c**) (**d**) (**e**)

Figure 2: Tile examples extracted from IHC 0+ (**a**), IHC 1+ (**b**), IHC 2+ (**c**), IHC 3+ (**d**), H&E (**e**) slides. Examples extracted from [7, 8, 9].

3 Data & Training Details

The dataset is composed of subsets of slides from two public datasets: the HER2 Scoring Contest (HER2SC) training set [7] and the TCGA-TCIA-BRCA (BRCA) collection [8, 9]. The HER2SC training set (with available labelling) comprises slides of 52 cases of invasive BCa stained with both IHC and H&E. The subset from the BRCA dataset includes 54 H&E slides. All slides have the same original resolution and are weakly annotated with HER2 status (negative/positive) and score (0+, 1+, 2+, 3+), obtained from the corresponding histopathological reports. The training and validation sets, used for the IHC model parameter tuning and optimization, have 40 and 12 slides, respectively, corresponding to 7591 tiles per class for training (30,364 tiles total) and 624 tiles per class for validation (2496 tiles total), to keep a class balance.

The hyperparameters used during training were empirically set to maximize performance. The CNN model for IHC tile scoring was randomly initialized and trained using the Adaptive Moment Estimation (Adam) optimizer (learning rate of 1×10^{-5}), to minimize a cross-entropy loss function, during 200 epochs, with mini-batches of 128 tiles. The soft-argmax used a parameter $\beta = 1000$. The aggregation MLP was also trained using the Adam optimizer, with learning rate of 10^{-5} for 150 epochs and mini-batches of 1 slide (consisting of soft-argmax scores of the respective 300 tiles), saving the best considering validation accuracy.

4 Results and Discussion

After training, the IHC model offered 76.8% accuracy. This indicates that the model was able to adequately discriminate against the IHC tiles between the four classes.

On the HER2SC test set, this method achieved a weighted accuracy of 83.3% and a F1-score of 86.7% (see Table 1). Despite the small size of this test set, the proposed method was able to correctly classify all positive slides and only misclassify one negative sample as positive. In this context, one might consider this a desirable behaviour, as false positives are less impactful than false negatives.

Table 1: H&E HER2 status classification results of the proposed method.

	Accuracy	**F1-score**	**Precision**	**Recall**
HER2SC	83.3%	86.7%	89.6%	87.5%
BRCA	53.8%	21.5%	81.2%	31.5%

When tested on the BRCA test set, this method achieved a weighted accuracy of 53.8% and a F1-score of 21.5% (see Table 1). The method retains the behaviour presented in HER2SC, preferring to err on the side of false positives than the alternative. On the other hand, the performance metrics on BRCA differ considerably from those obtained on HER2SC. While the method was trained on HER2SC data, which is expected similar to the test data, the slides of the BRCA have a greater extent of tissue, generating more tiles per image and impacting the distribution of the scores, which may influence the method's behaviour.

The other shortcomings of the method appear to be related to the invasive tumour segmentation and the tile scoring network, which could be improved with additional data and more accurate ground truth. With these additional efforts, the proposed method could offer robust weakly-supervised HER2 classification without IHC information.

5 Conclusions

In this work, a framework is proposed for the weakly supervised classification of HER2 overexpression status on H&E BCa slides. The proposed approach integrates a CNN trained for HER2 scoring of individual H&E tiles, initialized with the network parameters pre-trained with data from IHC images. The objective of this initialization is to transfer some domain knowledge to the final training. The individual scores are aggregated on a single prediction per slide, returning the HER2 status label.

The evaluation results in single-database (HER2SC) and cross-database (BRCA) settings show the potential of the proposed method in standard and more challenging situations, indicating that it is possible to accurately infer BCa HER2 status solely from H&E slides.

Despite these results, further efforts should be devoted to performance improvement. Firstly, the training of the tile HER2 scoring CNN and the aggregation MLP could be integrated into a single optimization process. On the other hand, the aggregation of individual scores could use tile locations to take spatial consistency into account. Finally, the knowledge embedded in the networks through the pre-trained parameters could be better seized if input H&E tiles could be previously converted into IHC, for example, using generative adversarial models.

Acknowledgements This work was partially funded by the Project "TAMI: Transparent Artificial Medical Intelligence" (NORTE-01-0247-FEDER-045905), co-financed by ERDF, European Regional Fund through the Operational Program for Competitiveness and Internationalisation (COMPETE 2020), the North Portugal Regional Operational Program (NORTE 2020) and by the Portuguese Foundation for Science and Technology (FCT), under the CMU-Portugal International Partnership, and also the FCT PhD grants "SFRH/BD/139108/2018", "SFRH/BD/137720/2018" and "SFRH/BD/06434/2020".

References

[1] American Cancer Society. Breast Cancer Facts & Figures 2017–2018. Available online: `http://bit.ly/acs_bcff_1718`.

[2] Gandomkar, Z.; Brennan, P.; Mello-Thoms, C. Computer-based image analysis in breast pathology. *J. Pathol. Inform.* **2016**, *7*.

[3] Veta, M.; Pluim, J.P.W.; van Diest, P.J.; Viergever, M.A. Breast Cancer Histopathology Image Analysis: A Review. *IEEE Trans. Biomed. Eng.* **2014**.

[4] American Society of Clinical Oncology (ASCO). Breast Cancer Guide. 2005–2020. Available online: `http://bit.ly/asco_bcg`.

[5] Rakha, E.A. *et al*. Updated UK Recommendations for HER2 assessment in breast cancer. *J. Clin. Pathol.* **2015**, *68*, 93–99.

[6] Wolff, A.C. *et al*. Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Focused Update. *J. Clin. Oncol.* **2018**, *36*, 2105–2122.

[7] Qaiser, T. *et al*. HER2 challenge contest: A detailed assessment of automated HER2 scoring algorithms in whole slide images of breast cancer tissues. *Histopathology* **2018**, *72*, 227–238.

[8] Clark, K. *et al*. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *J. Digit. Imaging* **2013**, *26*, 1045-1057.

[9] Lingle, W. *et al*. Radiology Data from The Cancer Genome Atlas Breast Invasive Carcinoma [TCGA-BRCA] collection. *Cancer Imaging Arch.* **2016**.

[10] Cruz-Roa, A. *et al*. High-throughput adaptive sampling for whole-slide histopathology image analysis (HASHI) via convolutional neural networks: Application to invasive breast cancer detection. *PLoS ONE* **2018**, *13*, 1–23.

[11] Cruz-Roa, A. *et al*. Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent. *Sci. Rep.* **2017**, *7*.

[12] Honari, S. *et al*. Improving landmark localization with semi-supervised learning. In CVPR, 19–21 June 2018; pp. 1546–1555.