Predicting Residual Cancer Burden in a triple negative breast cancer cohort

Peter Naylor, Joseph Boyd, Marick Lae, Fabien Reyal, Thomas Walter

Predicting Residual Cancer Burden in a triple negative breast cancer cohort

Peter Naylor[1,2,3], Joseph Boyd[1,2,3], Marick Laé[4], Fabien Reyal[5,6,7], and Thomas Walter[1,2,3]

[1]MINES ParisTech, PSL Research University, CBIO - Centre de Bioinformatique, 75272 Paris Cedex 06
[2]Institut Curie, 75248 Paris Cedex 05
[3]INSERM U900, 75248 Paris Cedex 05
[4]Service de Pathologie, Institut Curie, Paris, F-75248, France
[5]Residual Tumor & Response to Treatment Laboratory, RT2Lab, Translational Research Department, Institut Curie
[5]U932, Immunity and Cancer, INSERM, Institut Curie
[6]Department of Surgery, Institut Curie, Paris, F-75248, France

August 2, 2019

Abstract

Analysis and interpretation of stained histopathology sections is one of the main tools in cancer diagnosis and prognosis. In addition to the information which is typically extracted by trained pathologists, there is also information that is not yet exploited, simply because we do not yet understand the impact of all cellular and tissular features that could be predictive of outcome. In this paper, we address a question that can currently not be solved by pathologists: the prediction of treatment efficiency for Triple Negative Breast Cancer (TNBC) patients from biopsy data.

Keywords— B reast Cancer, Computer-aided detection and diagnosis, Deep Learning, Digital Pathology, Histopathology, Triple Negative Breast Cancer

1 Introduction

Among women in France breast cancer is the most common cancer and leading cause of cancer deaths with 18.2% of deaths among female cancer patients[7]. TNBC is a subtype of breast cancer with poor prognosis and limited treatment options. In TNBC, the malignant invasive cells do not contain receptors for estrogen (ER), progesterone (PR) or HER2 and can therefore not be treated with

hormone therapies or medications that work by blocking HER2. The treatment used is neoadjuvant chemotherapy, i.e. chemotherapy prior to surgery. Response to neoadjuvant chemotherapy varies among patients and can be quantified after surgery via a Residual Cancer Burden (RCB) score [15], which is determined in a pathologic examination after treatment and is based on the size of the primary tumour bed area, overall Cancer Cellularity (as percentage of area), percentage of Cancer that is *in situ*, the number of metastatic axillary lymph and the diameter of the largest metastasis in an axillary lymph. It has been shown that the RCB score correlates well with survival. An RCB score of 0 means that the patient achieved pathologic complete response (pCR), whereas an RCB score of 1 or above implies a higher level/stage of RCB, denoted as RCB-I, RCB-II or RCB-III. A grade of pCR or RCB-I is highly correlated with a good prognosis [15]. By definition, the RCB score is evaluated after analysis of the surgical specimen once the chemotherapy is completed.

Here, we aim at predicting the RCB score as a measure of treatment efficiency from a biopsy that is taken prior to the treatment. In particular we shall compare multiple strategies for prediction: a manual feature extraction model and 2 automatic feature extraction models, namely a two step method and a fully-supervised method. It will be interesting to see whether and to which extend such a prediction is possible; the work we present here can therefore be seen as a contribution to a long-term effort to understand the resistance to treatment.

2 Related work

With the advent of digital pathology, the availability of large annotated data sets and the progress related to deep learning, computational analysis of histopathology data has known an enormous increase in popularity. Not only the number of publications in this field is increasing dramatically, but also the heterogeneity of tasks that is addressed by such algorithms is ever increasing. One important task relates to image segmentation, such as segmentation of nuclei in tissue [14, 12], detection and segmentation of metastatic regions in whole slide images (WSI) [18], classification of regions into normal, benign, *in situ* or invasive [1]. Other approaches target automatic grading and prognosis [3, 12, 19].

Importantly, in order to reach a higher level of reproducibility, a fair comparison of computational methods and important resources of annotated data, large-scale challenges have been organized in this field: for metastasis segmentation [13], region classification [1] and proliferation prediction [16].

One important question in the field is how to encode the information in an entire whole slide image (WSI), that is typically of size 100000×100000 pixels. The common procedure is to extract tiles from the WSI, process these tiles and aggregate the information for the prediction task at hand [17, 19, 6]. In [19], the authors discriminate tiles with unsupervised learning and find relevant groups, these relevant tiles are then used to determine a patient's prognosis. In [6], the authors demonstrate that they only require patient level annotation to

yield comparable results to methods using a pixel wise annotation, where each metastasis was segmented. This type of weakly supervised learning has already been proposed in the past for cancer region segmentation in colon cancer [18] and provides also the rationale for our study.

Of note, the major body of approaches in this field aims at automatizing or supporting the work currently performed by pathologists. This is to be contrasted to methods where the predictive power of biological evidence in the data remains unclear for the prediction task as in the project we present in this article.

3 Data set

The dataset was generated at the Curie Institute and consists of annotated H&E stained histology images at 40× magnification. In this paper, we studied chemotherapy treatment response in a unpublished cohort of TNBC. Prior to chemotherapy a biopsy was sampled from the tumour region and pathological relevant features were extracted, these features are discussed in more details in section 4.1. Post treatment, an RCB score is derived. The data was quality checked and analysed by an expert histopathologist, a total of 122 histopathological slides were annotated. 56 of these are annotated pCR, 10 are RCB-I, 49 are RCB-II and 7 are RCB-III. As some classes are under-represented we decided to simplify the problem to a binary classification. We investigate two possibilities: (1) pCR (no residuum) vs RCB (some residuum) and (2) pCR-RCB-I vs RCB-II-III, which is clinically more relevant, as it is informative on patient's prognosis.

4 Methodology

For each of the following methods, we will denote by $(X_i, Y_i)_{i\in\{1,..,N\}}$ the data set described in 3. N denotes the number of patients in the study, $N = 122$. For patient i, X_i represents the set of associated input data, such as the feature vector extracted by a pathologist or a bag of features extracted automatically from the biopsy. The binary output variable Y_i reflects the RCB class as described in section 3. In order to have an unbiased estimation of the accuracy, we perform a 10-fold nested cross-validation where the inner-loop maximizes, over the hyper-parameter, a validation accuracy score. Once the hyper-parameters set, we retrain with the maximum amount of available data except for the left-out fold and report the accuracy over this fold. The final score is the average of the 10-fold nested cross-validation. Furthermore on the hyper-parameters, we tuned the learning rate that could be chosen from $[10^{-i}; i \in [1:4]]$ and the number of units in the fully connected layers that could be chosen from $[128, 512, 1024, 2048]$.

4.1 Manual feature extraction

Each biopsy was analysed by a pathologist with strong expertise in breast cancer. Concretely, several regions of interest are analysed and several variables reported: Mitotic index on 10 high power fields (x40), Elston Ellis grade, percentage of cancerous cells (including CIS), percentage of Tumour Infiltrating Lymphocytes (TIL) and the percentage of lymphocyte in stroma. These features are informative about cancer prognosis [2]. In this setting, X_i will represent a feature vector of size 5. We used traditional machine learning models such as Random Forest (RF) [4] with randomized hyper-parameter search. We report the best model in section 5.

4.2 Automatic feature extraction

Here, we seek a way of automatically encoding an entire WSI. The advantage of such a description with respect to manual analysis or computational approaches based on down sampling is the comprehensiveness of the representation. In addition, we hope to extract pieces of information complementary to what a manual annotator would typically look at.

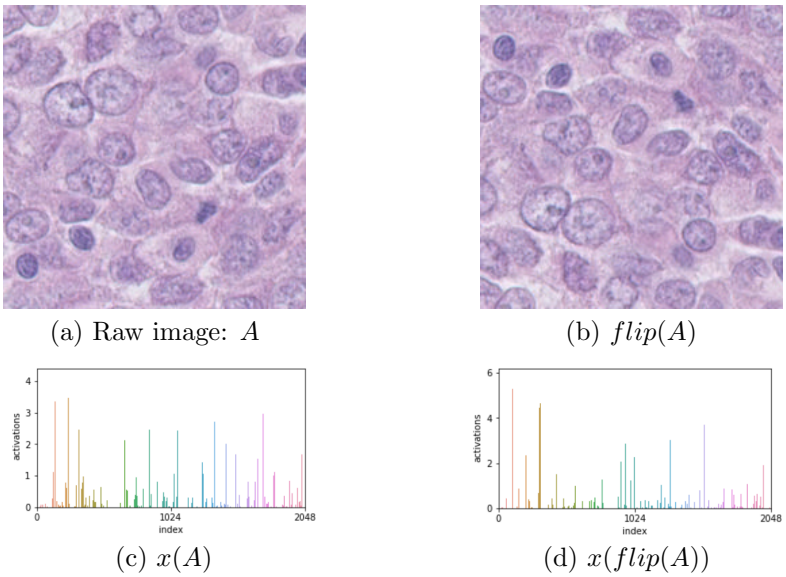
In the following methods, a WSI is represented by a bag of features. This mapping can be divided into 3 steps: 1) finding tissue areas in the WSI, 2) overlaying a grid on this tissue area and 3) encoding each tile of size 224×224 to a vector $x \in \mathbb{R}^p$ with a 152-layer Resnet [9] pre-trained on Imagenet [11], we have $p = 2048$. Thus each slide can be condensed into a matrix where each tile corresponds to one line of size p.

4.2.1 Rotational invariance

We notice undesirable sources of variation in the encoding: rotating an image can lead to very different representations, see Figure 1. Histopathology images clearly should have an encoding that is invariant to rotations and flips. It is not surprising that this invariance is not preserved with weights trained on Imagenet, as orientation often matters for natural images. In order to obtain a more robust representation of each sub-image we perform 6 augmentations for each image and aggregate the resulting representations via a pooling operator, average or maximum along the feature axis. We experiment and compare these 2 different pooling methods to one where no pooling is applied (None), see table 1.

4.2.2 A two step method

Here, we propose a two step model. In this model, we assume that a slide is made of several regions with varying importance for patient level prediction. We therefore hypothesized that if we can cluster the encoding vectors into k clusters, each slide can be represented by the percentages of tiles in each of the clusters. A similar approach has been reported in [19]. One technical issue is however the



(a) Raw image: A (b) $flip(A)$

(c) $x(A)$ (d) $x(flip(A))$

Figure 1: Rotational variance of the ResNet encoding: (a) and (b) show a tile and its flipped version, respectively. (c) and (d) show the corresponding encodings $x \in \mathbb{R}^P$. The color code indicates the index in the feature vector and allows for comparison between (c) and (d).

large number of tiles, making the application of unsupervised learning methods difficult if not impossible. This can be addressed by sub-sampling the data.

First, we reduce the dimensionality of each tile via a principal component analysis and keep 50 features (this is more than 99% of the explained variance). Then we sub-sample each slide to the same number of tiles. The two different sub-sampling strategies for each bag of features of size n_i are:

1. *uniform sampling*: we draw randomly a fixed number of feature vectors
2. *cluster-based down sampling*: we first cluster all feature vectors from one patient into $n_i/40$ clusters and then sample the same (small) number of feature vectors from each cluster so that the amount of feature vectors is constant across patients.

Finally, we pool all sampled feature vectors and perform clustering into k clusters on their union. By doing this, we can assign to each tile of each patient a cluster label $y_j^{(i)}$. Once each tile is clustered, we thus represent a WSI by the percentage of patches belonging to each of the k clusters. Hence, we represent a patient's biopsy by a vector $z^{(i)}$ of size k. We then use a random forest to classify each patient, see figure 2. We set $k = 4$ as this gave best results in our hands and this choice is in line with the recently published BACH 2018 challenge for WSI region annotation in breast cancer histology [1].

4.2.3 Fully supervised

Adopting a multiple instance learning framework allows us to adopt an end-to-end learning approach from the encoded tiles to the patient-level output. In this setting, $X_i = (x_j^{(i)})_{j=1...n_i}$ is a bag of features of variable size, i.e. each slide is represented by a variable number n_i of feature vectors $x_j^{(i)}$ each of which describes one tile. Instead of assigning to each $x_j^{(i)}$ one hard cluster label, as described in section 4.2.2, we map each $x_j^{(i)}$ to a low dimensional representation $y_j^{(i)} \in \mathbb{R}^K$. This can be seen as a generalization of a cluster assignment. We use a 1-dimensional convolution to learn the mapping. We then have to pool these representations $y_j^{(i)}$ for all j to reach a description $z^{(i)}$ of the entire slide. We feed this slide encoding to two fully connected layers with 512 units, see figure 3. For comparability with the previous approach, we set $K = 4$. If we set $K = 1$ and use the Weldon pooling [8] this model is essentially the one proposed by [6], our results with such a network were inconclusive.

Notes on the implementation: 1) We regularize the model with batch normalization, implying that the number of tiles per patient needs to be the same. For WSI, it is impracticable to follow an up-sampling strategy where each slide would be represented by the maximum number of tiles in the data set. We therefore use a down sampling and up sampling strategy to a fix amount. Up sampling: we add instances from the initial distribution. This has no effect later in the network as we aggregate the instances after the bottleneck. Down sampling: each time a slide requires down sampling, we randomly remove tiles. 2) During inference we do not perform any down sampling or up sampling. 3) We set the amount of samples in a bag to 5000. 4) We train the model with Adam optimizer [10] and use a binary cross entropy loss. 5) We use heavy regularization by setting a weight decay term to 0.05 and drop out to 0.5. 6) We repeat this 20 times per fold and retain the model that performs best on the validation data. 7) We choose this architecture to replicate the steps from our previous method. 8) We used Keras for the implementation [5].

We compare this model to a simpler version where we do not transform the features vector prior to averaging and use a RF for the classification, we denote this model as averaging the bag.

5 Results and discussion

In table 1 we display the performance of every combination of strategies and models described in section 4 on the two tasks described in section 3. These two tasks are very difficult, indeed even with the manual extraction of pathological relevant features it is difficult to correctly predict a patients class, reaching an accuracy of 57.0% on the whole data for task (1) and 59.8% for task (2). The results displayed in table 1 show similar performance to that of a pathologist. Therefore in terms of automatic feature extraction, the ResNet algorithm clearly catches some useful signal for predicting patient treatment response. In our

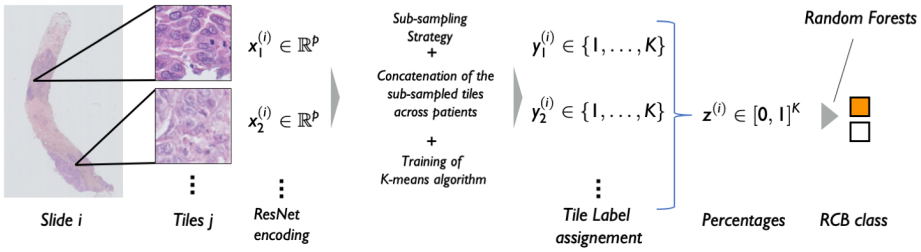


Figure 2: Two step method: ResNet features $x_j \in \mathbb{R}^P$ are extracted and used for clustering. Each tile j is therefore represented by the cluster label y_j and the slide is represented by the percentages of tiles in each of the clusters.

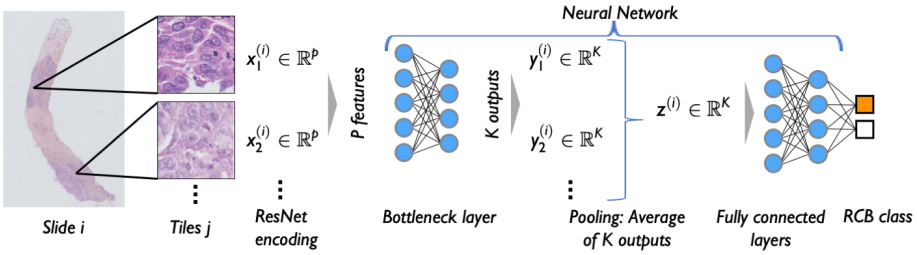


Figure 3: Fully supervised approach: the clustering approach in Figure 2 is replaced by a Neural Network.

2 step method, we notice a 5% improvement of the k-means sampling over a uniform distribution. This would suggest that random sub-sampling of tiles damages the performances. This suggests that the tiles relevant for this decision might be scarce. We also note the improvement achieved by augmenting the images prior to encoding in most cases. We recommend to use one of the aggregation strategies to reduce the noise when encoding the tiles. The neural network model was only able to learn something meaningful in one of these two augmentation settings. Finally, the bottleneck layer finds a suitable K-class representation of the WSI for classification. This setup clearly outperforms a more basic approach where the slide is represented by the average of the bag.

6 Conclusion

Here, we presented and compared several methods suited to encode entire WSI. In particular, we tackled the challenging question of treatment response prediction from biopsy images and show that automatic methods reach the performance of a pathologist. We also highlighted the most important design choices, and propose a strategy to fight rotational variance of the encodings, so far neglected in the field.

Model:		(1)	(2)
Manual:		57.0%	59.8%
2 step method:			
Uniform:			
Augmentations:	*None*:	52.5%	52.5%
	Max:	52.5%	53.2%
	Mean:	53.2%	53.2%
K-means:			
Augmentations:	*None*:	58.2%	56.6%
	Max:	58.2%	58.2%
	Mean:	59.9%	58.2%
Neural network:			
Augmentations:	*None*:	55.7%	50.8%
	Max:	59.8%	47.5%
	Mean:	48.4%	60.6%
Averaging the bag:			
Augmentations:	*None*:	48.3%	52.4%
	Max:	51.6%	54.1%
	Mean:	54.9%	54.9%

Table 1: Classification results in terms of accuracy, 10-fold nested cross validation. Column (1): pCR vs RCB and column (2): pCR-RCB-I vs RCB-II-III. *None*, *Max* and *Mean* refer to the augmentation methods for rotational invariance described in 4.2.1.

References

[1] Guilherme Aresta, Teresa Araújo, and Scotty et al. Kwok. Bach: Grand challenge on breast cancer histology images. *arXiv preprint arXiv:1808.04277*, 2018.

[2] Yuka Asano, Shinichiro Kashiwagi, and Wataru et al. Goto. Prediction of survival after neoadjuvant chemotherapy for breast cancer by evaluation of tumor-infiltrating lymphocytes and residual cancer burden. *BMC cancer*, 17(1):888, 2017.

[3] Laura E Boucheron, BS Manjunath, and Neal R Harvey. Use of imperfectly segmented nuclei in the classification of histopathology images of breast cancer. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 666–669. IEEE, 2010.

[4] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[5] François Chollet et al. Keras, 2015.

[6] Pierre Courtiol, Eric W Tramel, Marc Sanselme, and Gilles Wainrib. Classification and disease localization in histopathology using only global labels: A weakly-supervised approach. *arXiv preprint arXiv:1802.02212*, 2018.

[7] Institut National du Cancer. *Les cancers en France*. Springer-Verlag New York, Inc., 2017 edition, 2017.

[8] Thibaut Durand, Nicolas Thome, and Matthieu Cord. Weldon: Weakly supervised learning of deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4743–4752, 2016.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[12] Neeraj Kumar, Ruchika Verma, Ashish Arora, Abhay Kumar, Sanchit Gupta, Amit Sethi, and Peter H Gann. Convolutional neural networks for prostate cancer recurrence prediction. In *Medical Imaging 2017: Digital Pathology*, volume 10140, page 101400H. International Society for Optics and Photonics, 2017.

[13] Geert Litjens, Peter Bandi, and Babak et al. Ehteshami Bejnordi. 1399 h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset. *GigaScience*, 7(6):giy065, 2018.

[14] Peter Naylor, Marick Laé, Fabien Reyal, and Thomas Walter. Segmentation of nuclei in histopathology images by deep regression of the distance map. *IEEE Transactions on Medical Imaging*, 2018.

[15] W Fraser Symmans, Florentia Peintinger, and Christos et al. Hatzis. Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy. *Journal of Clinical Oncology*, 25(28):4414–4422, 2007.

[16] Mitko Veta, Yujing J Heng, and Nikolas et al. Stathonikos. Predicting breast tumor proliferation from whole-slide images: the tupac16 challenge. *arXiv preprint arXiv:1807.08284*, 2018.

[17] Yan Xu, Zhipeng Jia, Liang-Bo Wang, Yuqing Ai, Fang Zhang, Maode Lai, I Eric, and Chao Chang. Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. *BMC bioinformatics*, 18(1):281, 2017.

[18] Yan Xu, Jianwen Zhang, I Eric, Chao Chang, Maode Lai, and Zhuowen Tu. Context-constrained multiple instance learning for histopathology image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 623–630. Springer, 2012.

[19] Xinliang Zhu, Jiawen Yao, Feiyun Zhu, and Junzhou Huang. Wsisa: Making survival prediction from whole slide histopathological images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7234–7242, 2017.