A MULTI-TASK MULTIPLE INSTANCE LEARNING ALGORITHM TO ANALYZE LARGE WHOLE SLIDE IMAGES FROM BRIGHT CHALLENGE 2022

Niccolò Marini[1,2] *Marek Wodzinski*[1,3] *Manfredo Atzori*[1,4] *Henning Müller*[1,5]

[1] University of Applied Sciences Western Switzerland
Information Systems Institute, Sierre, Switzerland
[2] Department of Computer Science, University of Geneva, Switzerland
[3] AGH University of Science and Technology
Department of Measurement and Electronics, Krakow, Poland
[4] Department of Neurosciences, University of Padua, Italy
[5] Medical Faculty, University of Geneva, Switzerland

ABSTRACT

Malignant lesions in breast tissue specimen whole slide images (WSIs), may lead to a dangerous diagnosis, such as cancer. However, WSIs analysis is time-consuming and expensive, requiring the work of expert pathologists. This paper aims to present a method for the 2022 BRIGHT Challenge, that involves the analysis of breast WSIs. The organizers provided over 550 breast WSIs and over 3900 regions of interest (ROIs) to develop and validate methods for breast cancer images. The method presented in this work is based on a Multiple Instance Learning instance-based Convolutional Neural Network (CNN), allowing the combination of strongly-annotated data (from ROIs) and weakly-annotated data (from WSIs) via the optimization of a multi-task loss function. Furthermore, during the CNN training, the input patches are clustered and filtered according to their entropy, to reduce the non-informative content used to train the model. The CNN reaches an averaged F1-score = 0.63 $\pm$ 0.02 on the 3-class classification task and averaged F1-score = 0.39 $\pm$ 0.08 on the 6-class classification task, considering the validation partition; an averaged F1-score = 0.65 on the cancer risk classification task and averaged F1-score = 0.45 on the sub-typing cancer risk classification task, considering the best result achieved on the test partition. These results show that Multiple Instance Learning instance-based CNNs may represent a good resource to tackle this kind of problem.

Index Terms— Deep Learning, Image classification, BRIGHT, Breast cancer, Multiple Instance Learning

1. INTRODUCTION

Breast cancer image analysis is still a critical challenge, despite the recent advancements in the development of automatic methods for the analysis of images may reduce the efforts needed to diagnose it. The analysis of breast whole slide images (WSI) is still expensive and time-consuming, requiring the manual work of medical experts[1] (i.e. pathologists) and aims to identify lesions linked to dangerous conditions in tissue specimens [2], according to predefined grading systems. Grading systems aim to distinguish between different subtypes of cancer, to better identify the risk related to the disease, and plan the most effective treatment for the patient. However, some findings related to breast cancer subtypes include high tissue morphology variability and require the analysis of an expert pathologist [3]. This low intra- and inter-observer agreement [4] on the diagnosis limits the possibility to adopt the right treatment for the patient. Computational pathology is a domain involving the development of automatic algorithms for the analysis of WSIs [5]. Currently, convolutional neural networks (CNNs) are state-of-the-art algorithms for solving several tasks, such as classification. However, especially in computational pathology, CNN application still has to face several challenges. CNNs require large datasets, not always easy to be collected, to learn robust and invariant features to solve tasks. Furthermore, local annotations are needed to train CNN with fully-supervision, but they are expensive to be collected since a pathologist is required to produce them. New frameworks were developed to tackle these problems, such as weak-supervision [5] and self-supervision [6]. Weak-supervision involves algorithms, such as Multiple Instance Learning (MIL) [5, 7, 8, 9], trained using labels referring to the whole image (usually the most dangerous diagnosis), without any information about the regions of interest within the image. Currently, the MIL framework shows high performance in WSI analysis, allowing to build an image as a set of patches without any knowledge about their content. Self-supervision is a framework that allows learning features from unlabeled data, learning relationships between the data, such as similarity. Self-supervised algorithms (such as MoCo [10]) is usually adopted to pre-train a CNN before the training on a specific task, called downstream task. The

goal of the BRIGHT Challenge is to collect and rank automatic algorithms developed to classify breast cancer images, providing a dataset including 753 WSIs and over 4000 regions of interest with rare and atypical lesions, to alleviate the manual analysis made by pathologists. This paper aims to present a MIL CNN method to contribute to BRIGHT Challenge. The method optimizes a multi-task loss function: one term involves the classification of the WSIs, while the other is the classification of patches coming from the ROIs, to exploit the relatively small amount of data provided. Furthermore, non-informative patches are filtered, allowing to focus the learning process only on informative regions.

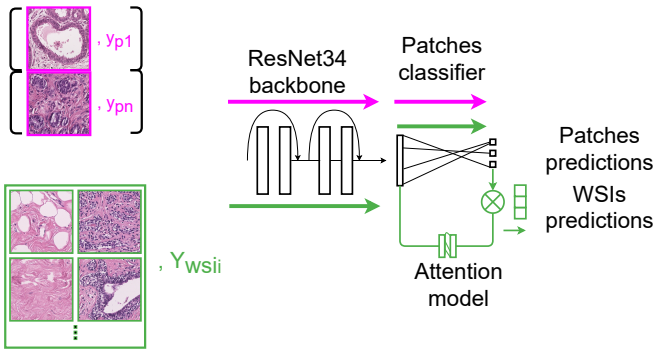

Fig. 1. Overview of the training schema for the multi-task Multiple Instance Learning CNN. The framework adopted is an instance-based MIL (the CNN produces predictions at both patch- and image-level).

2. METHODS

2.1. Data structure and pre-processing

The dataset includes 753 WSIs (423 for training, 80 for validation, 200 for testing) and 3989 regions of interest (ROIs, 3566 from training data and 489 from validation data), from BRACS dataset [2]. WSIs are gigapixel images (around $100'000^2$ pixels per image) globally annotated with a label representing a condition (Cancerous, Pre-cancerous, Non-Cancerous) or the breast morphology subtype (six subtypes). ROIs are smaller sections coming from the WSIs (around $2'000^2$ pixels per image). Due to ROIs small dimensions, the labels assigned to each section can be considered local annotations. Both WSIs and ROIs are split in patches of dimension 224x224 from magnification 10x, as shown in [11]. WSIs are split in a grid and only patches including tissue are extracted [12].

2.2. Multiple Instance Learning algorithm

The paper proposes an instance-based Multiple Instance Learning CNN combined with patch clustering to exploit the data (locally-annotated ROIs and globally-annotated WSIs), to classify breast cancer WSIs. An overview of the CNN is shown in Figure 1. The instance-based MIL CNN makes predictions on the single patches (y') and aggregates them through an attention-model [7] to make image-level predictions (Y'). The choice of adopting this framework is driven by the nature of data provided for the competition: WSIs with global annotations (i.e. the class) and small ROIs with local annotations. To exploit both the type of annotations, the optimization of the global loss function $Loss$ involves two terms: the first one, $Loss_{WSI}$, measures the error in the classification of WSIs, the second one, $Loss_{patches}$, measures the error in the classification of patches from the ROIs. A scalar coefficient λ is used to weigh the importance of the second term, as follows:

$$Loss_{WSI} = (\frac{1}{N}\sum_{n=1}^{N}\mathcal{L}_{CE}(Y_n, Y'_n) \quad (1)$$

$$Loss_{patches} = (\frac{1}{N}\sum_{n=1}^{N}\mathcal{L}_{CE}(y_n, y'_n) \quad (2)$$

$$Loss = Loss_{WSI} + \lambda * Loss_{patches} \quad (3)$$

WSIs are large images paired with global labels referring to a malignant condition that involves a small amount of tissue, even if most of the tissue is healthy or non-informative. To filter the non-informative tissue, a hierarchical patch clustering algorithm is adopted. The algorithm is applied to the patch embeddings (the number of levels varies depending on the number of patches included in the WSI). The CNN makes a cluster-level prediction and if the entropy of the cluster is over a threshold value, the patches are not considered in the WSI prediction, under the hypothesis that clusters with high entropy include heterogeneous patches and not a defined morphology.

2.3. Hyperparameters and Experimental setup

The CNN, for both the tasks, is implemented using PyTorch, with the same hyperparameters and the same training approach. The CNN has a ResNet34 backbone, including an additional intermediate layer (128 elements with Tanh as activation function) between the output of the ResNet and the classifier layer, and an attention network [7]. The CNN is pre-trained using MoCo v2, using a pipeline including patch rotation, flipping, color augmentation (involving the hue, the saturation, the contrast, and the RGB shift), Gaussian noise, elastic transformation, grid, and optical distortions. Furthermore, the CNN is pre-trained to learn features invariant to the H&E stain variability through a H&E-adversarial training [13]. For both MoCo and classification training, the hyperparameters are selected through a grid search algorithm, including the optimizer (Adam in both cases), the learning rate (10^{-4} for both cases), the weight decay (10^{-4} for both the cases) and the epochs (10 epochs for MoCo, 20 for classification). At the beginning of the training phase, until the

Table 1. Overview of the CNN results for the first task (3-class cancer risk WSI classification) on BRIGHT validation set. The results are evaluated considering the F1-score metric and compared with the baseline [11] provided by the organizers.

Method	**Noncancerous**	**Precancerous**	**Cancerous**	**Avg F1-score**
BRIGHT baseline	0.57 ± 0.05	0.43 ± 0.06	0.74 ± 0.03	0.58 ± 0.01
Our method	0.58 ± 0.03	0.51 ± 0.05	0.80 ± 0.03	0.63 ± 0.02

Table 2. Overview of the CNN results for the second task, 6-class subtyping WSI classification, on BRIGHT validation set. The subtypes are: Pathological Benign (PB), Usual Ductal Hyperplasia (UDH), Flat Epithelia Atypia (FEA), Atypical Ductal Hyperplasia (ADH), Ductal Carcinoma in Situ (DCIS), and Invasive Carcinoma (IC). The results are evaluated considering the F1-score metric and compared with the baseline [11] provided by the organizers.

Method	**PB**	**UDH**	**FEA**	**ADH**	**DCIS**	**IC**	**Avg F1-score**
BRIGHT baseline	0.43 ± 0.06	0.23 ± 0.08	0.20 ± 0.08	0.16 ± 0.08	0.41 ± 0.05	0.89 ± 0.01	0.39 ± 0.02
Our method	0.18 ± 0.10	0.36 ± 0.12	0.28 ± 0.14	0.22 ± 0.08	0.40 ± 0.23	0.87 ± 0.04	0.39 ± 0.08

first two epochs, only the images including less than 2'000 patches were used for training. The class unbalanced, for both the competition tasks, is tackled using a class-wise data augmentation approach. The model weights are saved only when the loss function is lower compared with the values reached in the previous epochs.

3. RESULTS

The proposed method outperforms the method provided as baseline [11] on the first classification task (3 classes), considering both the single class predictions and the global prediction, while, the performance reached on the second task (6 classes) is comparable to the one provided as the baseline. The classification performance is evaluated in terms of F1-score for each class and averaged F1-score for the cumulative results, on the validation partition, since the challenge organizers decided to not provide the labels for the test partition. Table 1 and Table 2 show the performance for respectively the 3-class cancer risk WSI classification and the 6-class subtyping WSI classification, reporting the average and the standard deviation of five models. The methods are also evaluated on a test partition including 200 WSIs (provided without ground truth). Considering the four submission made, the method obtains averaged F1-score = 0.611 ± 0.025 in the 3-class task and averaged F1-score = 0.427 ± 0.029. The highest score reached in the 3-class classification task is averaged F1-score = 0.65 (3rd highest score in the competition), while the highest score reached in the 6-class classification task is averaged F1-score = 0.45 (4th highest score).

4. DISCUSSION

The paper presents a contribution to the BRIGHT challenge with a MIL CNN that filters non-informative patches. The obtained results show good performance, outperforming the baseline proposed by organizers in the first task and reaching a result comparable with the baseline proposed by organizers. The performance of the CNN shows a good capability to classify the presence of Cancerous and Precancerous conditions, while the capability to classify NonCancerous class is similar to the baseline proposed by the organizers. This result may help pathologists since the analysis of Cancerous and Precancerous tissue morphologies require the work of experts pathologists. However, most of the samples analyzed in digital workflows include NonCancerous tissue. Therefore, the performance obtained on NonCancerous tissue may be considered a limitation for the application of this method in clinical settings. The CNN reaches a performance comparable to the baseline on the classification of tumor subtyping. Also in subtyping classification, the global CNN performance is hindered in the NonCancerous (PB and UDH) and PreCancerous (FEA and ADH) classification. In particular, while the results reached in PreCancerous subtypes are slightly higher than the baseline, the results reached in NonCancerous subtypes are lower, in particular regarding PB subtype. The result obtained on this task can be the consequence of aspects involving the method adopted. The first aspect involves the MIL instance-based framework. Multiple Instance Learning is a framework that requires large datasets to learn robust features from input data, while the training partition includes only 423 WSIs. The small dimension of the dataset influences particularly the performance on subtyping classification, where the CNN shows large standard deviation values. In particular, the instance-based framework produces an image-level prediction through the aggregation of the predictions on the single patches and was chosen considering the availability of ROIs. It is possible that the network overfitted on the patches coming from the ROIs, hindering the learning process of the attention network responsible for the aggregation. Another aspect involves the clustering algorithm, developed to discard non-informative patches during the training. The clustering algorithm, discarding non-informative patches, may have focused the learning process only on features linked to the malignant condition, explaining the good but not perfect results obtained on

PB in the subtype classification task. Furthermore, the NonCancerous class does not include morphologies that can lead to malignant conditions, but healthy or non-informative tissue, so the NonCancerous class (and its subtypes) represents the absence of the other classes (PreCancerous and Cancerous). On the other hand, the clustering algorithm, tuned with fixed thresholds, may remove a too large number of patches, including informative ones, contributing to increase the noise during the training, in addition to the weak labels, and forcing the network to not identify the correct relationships between patches. The high standard deviations obtained in the second task may confirm this lack of robustness in the predictions of the model.

5. ACKNOWLEDGMENTS AND COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by the BRIGHT challenge. Ethical approval was not required as confirmed by the license attached with the open access data. The authors declare no conflict of interest. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825292.

6. REFERENCES

[1] Elizabeth A Krupinski, Anna R Graham, and Ronald S Weinstein, "Characterizing the development of visual search expertise in pathology residents viewing whole slide images," *Human pathology*, vol. 44, no. 3, pp. 357–364, 2013.

[2] Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniel Riccio, Giosuè Scognamiglio, Guillaume Jaume, Giuseppe De Pietro, Maurizio Di Bonito, Antonio Foncubierta, Gerardo Botti, et al., "Bracs: A dataset for breast carcinoma subtyping in h&e histology images," *arXiv preprint arXiv:2111.04740*, 2021.

[3] Kimberly H Allison, Lisa M Reisch, Patricia A Carney, Donald L Weaver, Stuart J Schnitt, Frances P O'Malley, Berta M Geller, and Joann G Elmore, "Understanding diagnostic variability in breast pathology: lessons learned from an expert consensus review panel," *Histopathology*, vol. 65, no. 2, pp. 240–251, 2014.

[4] Joann G. Elmore, Gary M. Longton, Patricia A. Carney, Berta M. Geller, Tracy Onega, Anna N. A. Tosteson, Heidi D. Nelson, Margaret S. Pepe, Kimberly H. Allison, Stuart J. Schnitt, Frances P. O'Malley, and Donald L. Weaver, "Diagnostic concordance among pathologists interpreting breast biopsy specimens," *JAMA*, vol. 313, no. 11, pp. 1122–1132, 03 2015.

[5] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature medicine*, vol. 25, no. 8, pp. 1301–1309, 2019.

[6] Chetan L Srinidhi, Seung Wook Kim, Fu-Der Chen, and Anne L Martel, "Self-supervised driven consistency training for annotation efficient histopathology image analysis," *Medical Image Analysis*, vol. 75, pp. 102256, 2022.

[7] Maximilian Ilse, Jakub Tomczak, and Max Welling, "Attention-based deep multiple instance learning," in *International conference on machine learning*. PMLR, 2018, pp. 2127–2136.

[8] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature biomedical engineering*, vol. 5, no. 6, pp. 555–570, 2021.

[9] PJ Sudharshan, Caroline Petitjean, Fabio Spanhol, Luiz Eduardo Oliveira, Laurent Heutte, and Paul Honeine, "Multiple instance learning for histopathological breast cancer image classification," *Expert Systems with Applications*, vol. 117, pp. 103–111, 2019.

[10] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[11] Nadia Brancati, Giuseppe De Pietro, Daniel Riccio, and Maria Frucci, "Gigapixel histopathological image analysis using attention-based neural networks," *IEEE Access*, vol. 9, pp. 87552–87562, 2021.

[12] N Marini, S Otálora, D Podareanu, M van Rijthoven, J van der Laak, F Ciompi, H Müller, and M Atzori, "Multi_scale_tools: a python library to exploit multiscale whole slide images," *Frontiers in Computer Science*, vol. 68, 2021.

[13] Niccolò Marini, Manfredo Atzori, Sebastian Otálora, Stephane Marchand-Maillet, and Henning Müller, "H&e-adversarial network: A convolutional neural network to learn stain-invariant features through hematoxylin & eosin regression," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021, pp. 601–610.