Multi-Scale Multiple Instance Learning for the Classification of Histopathological Images with Global Annotations

First Author[1 [0000−1111−2222−3333]], Second Author[2,3 [1111−2222−3333−4444]], and Third Author[2 [2222−−3333−4444−5555]]

Anonymous

Abstract. Whole slide images (WSIs) are often provided with global annotations in the form of pathology reports. Local annotations are less frequently available, as obtaining them is time consuming. Global annotations do not include information about the regions of interest or the magnification levels used for the diagnosis. This fact can limit the training of machine learning models, as WSIs are usually very large and the part mentioned in the diagnosis can be very small. This paper presents a Multi-Scale Multiple Instance Learning (MSMIL) method, allowing to better exploit data paired with global labels, without local annotations and to combine contextual and detailed information identified at several magnification levels. The method is based on a MIL framework, to deal with the absence of local annotations and combines images from several magnification levels to deal with the absence of the magnification levels used for the diagnosis. The model produces a global prediction and a prediction for each magnification level used. MSMIL is evaluated on colon cancer images for binary and multilabel classification. MSMIL shows an improvement in performance above the single scale MIL and a global prediction multi-scale MIL, demonstrating that MSMIL can help to better deal with global labels targeting full and multi-scale images.

Keywords: Multi-Scale Multiple Instance Learning · Multiple Instance Learning · Multi-scale approach · Computational pathology.

1 Introduction

Histopathology is the gold standard for diagnosing many diseases like cancer [1]. Computational pathology involves the automatic analysis of digital histopathology images, usually in the form of whole slide images (WSIs). The WSIs include several magnification levels of the samples, since tissue patterns and morphology vary depending on the magnification at which they are viewed. Low magnification levels (5x) allow the visualization of glands, while higher magnification levels (20x-40x) allow the visualization of single cells. Training machine learning algorithms for the automatic analysis of digital pathology images is still an open challenge [6], also due to the limited availability of large datasets with local annotations.Convolutional Neural Network (CNNs) are currently the state-of-the-art

for computational pathology tasks such as classification of WSIs [21]. CNNs usually require many locally (pixel-wise) annotated samples to train models effectively [13]. Local annotations are not always available, as they are an expensive and time-consuming process that usually requires the involvement of pathologists. Most public datasets [8] do not include local annotations but many are paired with medical reports, high-level text descriptions of the image content including information used for the diagnosis. This information can be used as a global (weak) label for the image. This kind of label is inherently noisy [12]: the label refers to the whole image and it does not include any information regarding the regions of interest used for performing the diagnosis. The labels also do not include any information about the magnification levels used for the diagnosis.

Recently, new methods to face the lack of local annotations were proposed, such as Multiple Instance Learning (MIL). Regarding the lack of information about the magnification levels used, approaches to combine multi-scale images in CNN training were used. Few studies target the combination of both approaches. MIL [10, 5, 17, 18, 20, 22] includes weakly-supervised algorithms that allow facing the lack of information regarding the regions of interest. Histopathology image classification can be formulated as a MIL problem, where a WSI represents a bag X_n that includes P patches and the information available on the data regards the entire WSI. Approaches to combine multi-scale images in CNN training [10, 11, 3, 23, 15, 19] allow to face the lack of information regarding the magnification levels involved in the diagnosis, combining contextual and detailed information identified at several magnification level. The approaches can involve architectures where each magnification has its own branch to extract and combine features [10, 11, 23], U-Net based networks [3, 19] and CNNs where the convolution layers include multiple receptive fields [16, 15]. Few and only recent approaches combine MIL and multi-scale images, such as [10], where the authors present a Multi-Scale Multiple Instance Learning (MSMIL) CNN to classify benign vs. malignant lymphoma. The CNN combines features from multi-scale patches in a MIL framework to obtain a global prediction for the WSI. The model shows a performance improvement over a CNN trained with patches from a single magnification level. The model does not provide outcomes at single magnification levels, different from what pathologists concretely do. Pathologists usually analyze the contextual information of the tissue at low magnification levels, identifying regions of interest and then zooming through them to analyze the tissue details and to confirm the disease findings at lower levels. The global diagnosis is the result of the combination of the contextual and detailed information identified at several magnification level. The MSMIL method described in this paper allows facing the lack of pixel-wise annotations and different spatial resolutions in CNN training, producing multiple predictions. The MSMIL CNN has multiple scale branches as input (one for each magnification level) and produces multiple predictions as output (one for each magnification level and a global prediction combining several levels). The multiple outputs of the model allow to better optimize the entire model and take advantage of the combination of contextual

and detailed information, since the global prediction influences and is influenced by the single-scale predictions like in a diagnostic process.

The method proposed in this paper is applied to the binary and multilabel classification of colon cancer (colorectal cancer), the fourth most commonly diagnosed cancer in the world [2]. The diagnosis of the disease involves the detection of cancerous polyps [9], small agglomerations of cells, located on the colon border and the detection of glands. These tissue structures are usually identified combining the visualization of low and medium magnifications.The dataset analyzed in this article includes the corresponding global diagnosis. The diagnosis can include one or several colon tissue findings, among four classes: cancer, high-grade dysplasia (hgd), low-grade dysplasia (lgd) and hyperplastic polyp. The proposed MSMIL method outperforms both a Single-Scale Multiple Instance Learning method and a MSMIL method in binary and multilabel problems producing only global predictions in colon image classification.

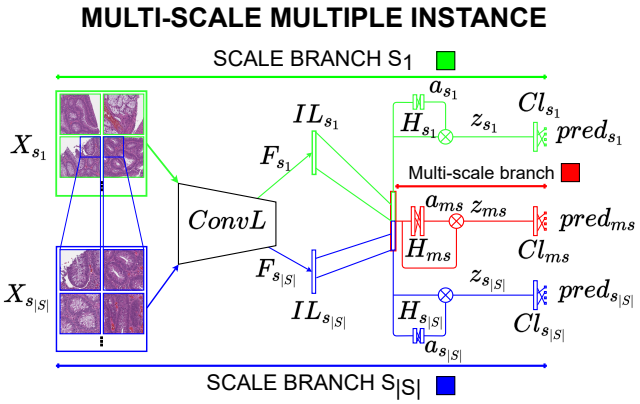

Fig. 1. Overview of the MSMIL model. The magnification levels are noted as s, the combined magnification levels as ms. X_s is a bag. ConvL is the convolutional layer block (shared among the branches). F_s is the feature vector, IL_s the intermediate fully-connected layer, H_s the embedding vector, z_s the output of the attention network, z_s the output of the attention network. Cl_s is the classifier, $pred_s$ the class prediction.

2 Methods

This paper proposes a MSMIL CNN to classify colon cancer WSIs. The method is based on CNNs that combine multi-scale images adopting a MIL framework.Figure 1 shows an overview of the CNN architecture. The magnification levels are noted as $s \in$ S ($|S|$ representing the number of magnification levels available). The CNN includes multiple scale branches ($|S|$ branches, $\{s_1, \ldots, s_{|S|}\}$, one for each magnification level as input) and produces $|S|$+1 predictions ($|S|$

single-scale predictions and one multi-scale prediction) as output. Each scale branch receives as input a WSI X_{ns}, the corresponding label Y_n and produces a prediction $pred_s$, for the corresponding magnification level s. Each scale branch includes convolutional layers, fully-connected layers, attention pooling layers and a classifier. The convolutional layers (ConvL) are used to extract the features (F_s). The fully-connected layers include an intermediate layer (IL_s), that produces smaller feature embeddings H_s from F_s, composed of the patch embeddings $\{h_p\}_s$ ($p \in$ P, $|P|$ representing the number of patches within a WSI). The attention pooling layer [17] aggregates the embeddings into a new array z_s, using an attention neural network (w_s and V_s are parameters of the network) that learns a function to weight (a_s are the attention weights for each class) the embeddings and produces and aggregated embedding $z_s = a_s \otimes H_s$.

$$z_s = (\sum_{p=1}^{P} a_p h_{p_s}) \tag{1}$$

$$a_p = \frac{\exp(w_s^T \tanh(V_s h_{p_s}))}{\sum_{j=1}^{P} \exp(w_s^T \tanh(V_s h_{j_s}))} \tag{2}$$

The classifier receives as input z_s and outputs the class prediction ($pred_s$), for a fixed magnification level. Each branch is trained to optimize a Binary-Cross entropy loss function. The CNN also includes a multi-scale branch that produces a multi-scale prediction by aggregating features from several scale branches. Multi-scale concatenated embedding ($h_{ms} = h_0, h_1, ...h_S$) feeds the multi-scale branch and another attention network ($a_m s$ as attention weights), producing multi-scale aggregated embeddings $z_m s = a_m s \otimes h_{ms}$. The embeddings are used to feed a classifier ($Cl_m s$) that outputs the multi-scale global prediction $pred_{ms}$. The multi-scale branch is trained to optimize a loss function (binary-cross entropy). The optimization process of the network involves a loss function with multiple terms. The terms in the equation are the multi-scale loss function (weighted with α) and the sum of the single-scale loss functions (weighted with β). This optimization leads to better performance also in the single-scale branches that benefit from the multi-scale features.

$$Loss = \alpha * Loss_{ms} + \beta * (\sum_{i=1}^{n} Loss_s) \tag{3}$$

3 Experiments

Dataset The MSMIL method is trained and evaluated on histopathology images of colon polyps acquired during colonoscopy. The colon dataset is from the ANONYMOUS SOURCE and is acquired with ethics approval. It includes 1478 WSIs from 947 patients, scanned with an Aperio and a 3DHistech scanners and stained with Hematoxylin and Eosin (H&E). All images include a global diagnosis of the images provided by a pathologist and a small subset comes with

pixel-wise annotations. The diagnosis includes one or more classes among: cancer, high-grade dysplasia, low-grade dysplasia and hyperplastic polyp. The WSIs are analyzed at 5-10x magnification, since pathologists recognize these classes at low to medium magnifications.The dataset is split into three partitions: training (1159 WSIs), validation (177 WSIs) and testing (142 WSIs pixel-wise annotated). The split is made considering the class distribution and that all images from a patient are included in the same partition.

Pre-processing The image pre-processing involves the image splitting into a bag X_n of patches $\{x_p\}$ and the linking between the instances from different magnification levels. A WSI is split into a grid of patches (only the ones including tissue are included) for each magnification. Patches are resized to 224x224 pixels after the extraction, regardless of the magnification level, to fit the pre-trained CNN architecture. The bags used to train the MSMIL model include patches from several magnifications: the i-th patch at a lower magnification includes the j-th patch within the bag at higher magnification. Considering that bags with patches from lower magnification include fewer patches than bags with patches from higher magnification, the i-th patch at lower magnification can be linked with more patches at higher magnification level.

Experimental setup The MSMIL and Multiple Instance single-scale CNNs have the same backbone architecture and are trained multiple times using the same strategy to set the hyperparameters to avoid overfitting and to face the class imbalance. The backbone architecture is a ResNet34 (pre-trained on ImageNet), used as a feature extractor. It produces feature vectors of size 512 for each input patch. Each model is trained five times to limit the non-deterministic effect of the stochastic gradient descent used to optimize the model using the chosen hyperparameters. The average and standard deviation of the models are reported. The hyperparameters are chosen with a grid search [7], aimed at finding the optimal configuration of the CNN hyperparameters (i.e. the configuration that allows the CNN to have the lowest loss function on the validation partition data). The hyperparameters involved in the grid search are the number of epochs (five epochs), the optimizer (Adam), the learning rate (10^{-3}), the decay rate (10^{-4}), the number of nodes within the intermediate fully-connected layers (128) and the value of α and β of the loss function (α=1 and β=1). Overfitting and class imbalance are limited adopting a class-wise data augmentation method that uses three operations: rotations, flipping and colour augmentation. The augmentation is implemented with the Albumentations library [4].

4 Results

MSMIL is evaluated considering the predictions of both the single-scale and multi-scale branches (global prediction), comparing the performance with two other MIL methods on binary and multilabel classification problems. MSMIL is compared with a Single-Scale MIL (SSMIL) and with a baseline MSMIL method

Table 1. Performance of MSMIL in the binary problem compared with SSMIL and baseline MSMIL using balanced accuracy, κ and F1-score.

MAGNIFICATION	balanced accuracy	κ	F1
5x (SSMIL)	0.877 ± 0.016	0.750 ± 0.036	0.874 ± 0.019
10x (SSMIL)	0.853 ± 0.039	0.711 ± 0.073	0.853 ± 0.039
MSMIL baseline [10]	0.881 ± 0.005	0.760 ± 0.010	0.879 ± 0.005
MSMIL global prediction	0.896 ± 0.010	0.791 ± 0.027	0.895 ± 0.010
MSMIL 5x branch prediction	**0.904 ± 0.010**	**0.808 ± 0.021**	**0.903 ± 0.010**
MSMIL 10x branch prediction	0.878 ± 0.018	0.752 ± 0.039	0.875 ± 0.020

Table 2. Performance of MSMIL for the multilabel problem compared with SSMIL and baseline MSMIL using macro accuracy, macro precision and macro recall.

MAGNIFICATION	macro accuracy	macro precision	macro recall
5x (SSMIL)	0.748 ± 0.067	0.714 ± 0.072	0.683 ± 0.125
10x (SSMIL)	0.821 ± 0.007	0.732 ± 0.068	0.710 ± 0.118
MSMIL baseline [10]	0.828 ± 0.013	0.734 ± 0.047	**0.74 ± 0.076**
MSMIL global prediction	0.835 ± 0.006	0.738 ± 0.035	0.691 ± 0.046
MSMIL 5x branch prediction	0.830 ± 0.014	0.775 ± 0.052	0.650 ± 0.068
MSMIL 10x branch prediction	**0.844 ± 0.012**	**0.788 ± 0.058**	0.683 ± 0.023

(only a global prediction), based on [10]. The baseline MSMIL CNN produces only a global WSI prediction. The implementation of the method includes colour augmentation instead of the domain adversarial network proposed by the authors to address colour variability and have a better comparison.

The binary problem involves the classification of high-risk classes (cancer and high-grade dysplasia) and low-risk classes (low-grade dysplasia and hyperplastic polyps). The performance is evaluated using balanced accuracy, Cohen's κ [14] and the F1 score. Table 1 summarizes the results. The CNN trained with the MSMIL method shows higher performance (for all metrics) in the binary WSI classification, compared with the SSMIL method and with the baseline MSMIL. The MSMIL single-scale branch trained with patches from 5x reaches the highest performance in all the metrics, even though it is comparable with the multi-scale prediction performance. The multilabel problem involves the classification of the four classes: cancer, high-grade dysplasia, low-grade dysplasia and hyperplastic polyps. The performance is evaluated using macro accuracy, macro precision and macro recall. Table 2 summarizes the results obtained. The CNNs trained with MISSL reaches the highest performance (for single-scale predictions and for the global prediction) in macro accuracy and macro precision, while it is outperformed in macro recall performance by the MSMIL baseline and by the SSMIL trained with patches from 10x. The performance of the MSMIL scale branch trained with patches from 10x obtains the highest performance in macro accuracy and macro precision.

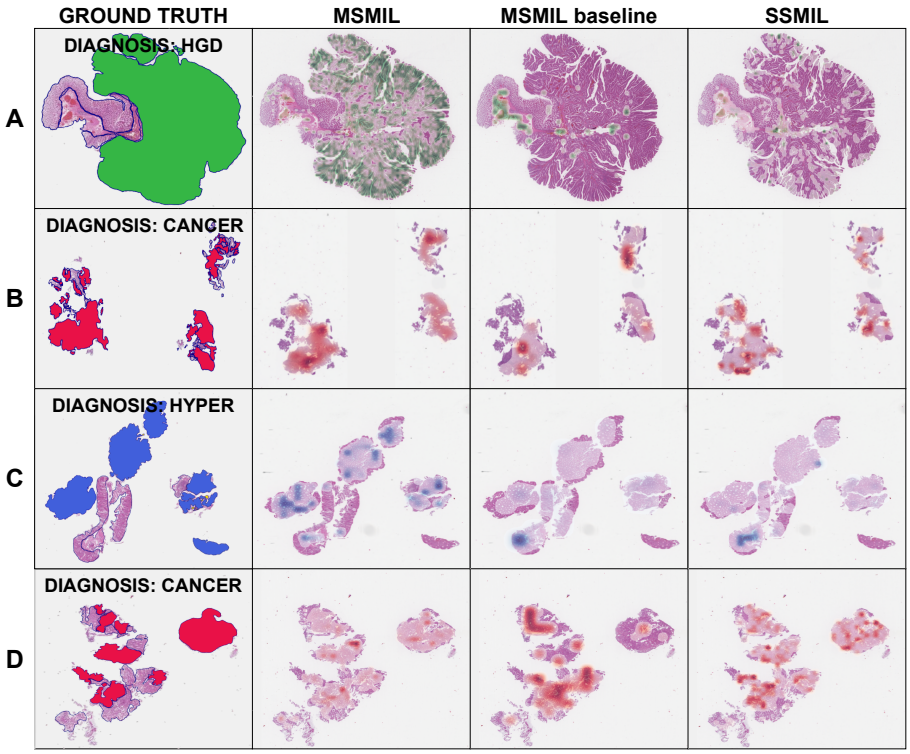

Fig. 2. Attention maps of MSMIL, the MSMIL baseline and SSMIL compared with pixel-wise annotations: cancer (red), hgd (green), lgd (yellow), hyperplastic polyp (blue), normal tissue (orange). In rows 1-3, MSMIL has best results, while in the last row MSMIL does not fully highlight the relevant areas.

5 Discussion

The results obtained show that the MSMIL CNN benefits of the multiple predictions, obtaining higher performance for most of the considered evaluation metrics compared with a SSMIL and a baseline MSMIL producing only a global prediction. Combining images from several magnification levels allows the model to focus on different details and combine both contextual and detailed information leading to the diagnosis. Figure 2 shows pixel-wise annotations made by a pathologist and attention heatmaps of MSMIL, baseline MSMIL and SSMIL in multilabel problem.In the top three rows, the attention maps produced by MSMIL correspond better to the pixel-wise annotations. In the last row the MSMIL baseline and SSMIL produce better attention maps. With multi-scale images as input and multiple predictions as output the models produce attentions maps focused on larger portions of the images, as shown in column MSMIL of Figure 2. This can be explained considering the multi-scale input images and the training optimization of MSMIL that allow the model to have a more de-

tailed feature representation. In the proposed MSMIL method the multi-scale loss function and the loss functions for each magnification level are optimized. In this way, updates of the parameters within a single-scale branch are influenced not only by the backpropagation but also by the other branches, since the features are combined in the multi-scale branch. Thus, the gradients are backpropagated into both the multi-scale and the single-scale branches, influencing the predictions and the branch attention weights.The results obtained show that for the binary and multilabel classification tasks the MSMIL CNN outperforms the single-scale CNN for most of the considered evaluation metrics and that all the scale branches benefit from the training with multi-scale images. In the binary problem, the multi-scale CNN shows higher performance than the single-scale CNN in all the metrics tested. In the multilabel problem, the presented MSMIL method shows higher macro accuracy and macro precision than the ones obtained by the single-scale CNN and by the baseline MSMIL. This result means that the model produces more accurate predictions and fewer false positives. However, the predictions of the single-scale CNN show higher recall, meaning that the MSMIL CNNs produce also more false negatives, while the single-scale CNNs produce fewer false negatives. This can be explained qualitatively evaluated considering the attention heatmaps in Figure 2. MSMIL has attention on larger regions and it is possible that it produces more false negatives regions and less conservative predictions that lead to more false negatives. The SSMIL and the baseline MSMIL are more conservative in the attention, focusing usually only on small regions and then producing few false negatives.

6 Conclusions

This paper introduces a novel MSMIL CNN to classify WSIs. The approach allows combining contextual and detailed information from multiple magnification levels. It has multiple scale branches as input and produces multiple single-scale and one multi-scale prediction. The MSMIL outperforms a SSMIL CNN and a MSMIL CNN that produces only a global prediction in colon WSI classification, both in binary and in multilabel classification. We plan to test MSMIL on additional data, other organs and with a larger number of scales. The code with the model pre-processing and implementation will be made publicly available on Github on publication, allowing reuse ans reproduxibility.

References

1. Aeffner, F., Wilson, K., Martin, N.T., Black, J.C., Hendriks, C.L.L., Bolon, B., Rudmann, D.G., Gianani, R., Koegler, S.R., Krueger, J., et al.: The gold standard paradox in digital image analysis: manual versus automated scoring as ground truth. Archives of pathology & laboratory medicine **141**(9), 1267–1275 (2017)
2. Benson, A.B., Venook, A.P., Al-Hawary, M.M., Cederquist, L., Chen, Y.J., Ciombor, K.K., Cohen, S., Cooper, H.S., Deming, D., Engstrom, P.F., et al.: Nccn guidelines insights: colon cancer, version 2.2018. Journal of the National Comprehensive Cancer Network **16**(4), 359–369 (2018)

3. Bozkurt, A., Kose, K., Alessi-Fox, C., Gill, M., Dy, J., Brooks, D., Rajadhyaksha, M.: A multiresolution convolutional neural network with partial label training for annotating reflectance confocal microscopy images of skin. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 292–299. Springer (2018)
4. Buslaev, A., Parinov, A., Khvedchenya, E., Iglovikov, V.I., Kalinin, A.A.: Albumentations: fast and flexible image augmentations. ArXiv e-prints (2018)
5. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Silva, V.W.K., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature medicine **25**(8), 1301–1309 (2019)
6. Cheplygina, V., de Bruijne, M., Pluim, J.P.: Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Medical image analysis **54**, 280–296 (2019)
7. Chicco, D.: Ten quick tips for machine learning in computational biology. BioData mining **10**(1), 35 (2017)
8. Courtiol, P., Tramel, E.W., Sanselme, M., Wainrib, G.: Classification and disease localization in histopathology using only global labels: A weakly-supervised approach. arXiv preprint arXiv:1802.02212 (2018)
9. Ferlitsch, M., Moss, A., Hassan, C., Bhandari, P., Dumonceau, J.M., Paspatis, G., Jover, R., Langner, C., Bronzwaer, M., Nalankilli, K., et al.: Colorectal polypectomy and endoscopic mucosal resection (emr): European society of gastrointestinal endoscopy (esge) clinical guideline. Endoscopy **49**(3), 270–297 (2017)
10. Hashimoto, N., Fukushima, D., Koga, R., Takagi, Y., Ko, K., Kohno, K., Nakaguro, M., Nakamura, S., Hontani, H., Takeuchi, I.: Multi-scale domain-adversarial multiple-instance cnn for cancer subtype classification with unannotated histopathological images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3852–3861 (2020)
11. Jain, M.S., Massoud, T.F.: Predicting tumour mutational burden from histopathological images using multiscale deep learning. Nature Machine Intelligence **2**(6), 356–362 (2020)
12. Karimi, D., Dou, H., Warfield, S.K., Gholipour, A.: Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. Medical Image Analysis **65**, 101759 (2020)
13. Komura, D., Ishikawa, S.: Machine learning methods for histopathological image analysis. Computational and structural biotechnology journal **16**, 34–42 (2018)
14. Kvålseth, T.O.: Note on cohen's kappa. Psychological reports **65**(1), 223–226 (1989)
15. Lai, Z., Deng, H.: Multiscale high-level feature fusion for histopathological image classification. Computational and mathematical methods in medicine **2017** (2017)
16. Li, S., Liu, Y., Sui, X., Chen, C., Tjio, G., Ting, D.S.W., Goh, R.S.M.: Multi-instance multi-scale cnn for medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 531–539. Springer (2019)
17. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data efficient and weakly supervised computational pathology on whole slide images. arXiv preprint arXiv:2004.09666 (2020)
18. Mercan, C., Aksoy, S., Mercan, E., Shapiro, L.G., Weaver, D.L., Elmore, J.G.: Multi-instance multi-label learning for multi-class classification of whole slide breast histopathology images. IEEE transactions on medical imaging **37**(1), 316–325 (2017)

19. van Rijthoven, M., Balkenhol, M., Siliņa, K., van der Laak, J., Ciompi, F.: Hooknet: multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images. Medical Image Analysis **68**, 101890 (2021)
20. Sudharshan, P., Petitjean, C., Spanhol, F., Oliveira, L.E., Heutte, L., Honeine, P.: Multiple instance learning for histopathological breast cancer image classification. Expert Systems with Applications **117**, 103–111 (2019)
21. Jimenez-del Toro, O., Otálora, S., Andersson, M., Eurén, K., Hedlund, M., Rousson, M., Müller, H., Atzori, M.: Analysis of histopathology images: From traditional machine learning to deep learning. In: Biomedical Texture Analysis, pp. 281–314. Elsevier (2017)
22. Wang, Y., Li, J., Metze, F.: A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 31–35. IEEE (2019)
23. Yang, Z., Ran, L., Zhang, S., Xia, Y., Zhang, Y.: Ems-net: Ensemble of multiscale convolutional neural networks for classification of breast cancer histology images. Neurocomputing **366**, 46–53 (2019)