Weakly Supervised Multiple Instance Learning Histopathological Tumor Segmentation

Marvin Lerousseau[1,2(✉)], Maria Vakalopoulou[1,2], Marion Classe[1,3],
Julien Adam[3], Enzo Battistella[1,2], Alexandre Carré[1], Théo Estienne[1,2],
Théophraste Henry[1], Eric Deutsch[1], and Nikos Paragios[4]

[1] Paris-Saclay University, Gustave Roussy, Inserm, 94800 Villejuif, France
[2] Paris-Saclay University, CentraleSupélec, 91190 Gif-sur-Yvette, France
`marvin.lerousseau@centralesupelec.fr`
[3] Gustave Roussy, Pathology Department, 94800 Villejuif, France
[4] TheraPanacea, 75014 Paris, France

Abstract. Histopathological image segmentation is a challenging and important topic in medical imaging with tremendous potential impact in clinical practice. State of the art methods rely on hand-crafted annotations which hinder clinical translation since histology suffers from significant variations between cancer phenotypes. In this paper, we propose a weakly supervised framework for whole slide imaging segmentation that relies on standard clinical annotations, available in most medical systems. In particular, we exploit a multiple instance learning scheme for training models. The proposed framework has been evaluated on multi-locations and multi-centric public data from The Cancer Genome Atlas and the PatchCamelyon dataset. Promising results when compared with experts' annotations demonstrate the potentials of the presented approach. The complete framework, including 6481 generated tumor maps and data processing, is available at https://github.com/marvinler/tcga_segmentation.

Keywords: Weakly supervised learning · Histopathological segmentation · Multiple instance learning · Tumor segmentation

1 Introduction

In digital pathology, whole slide images (WSI) are considered the golden standard for primary diagnosis [13,16]. The use of computer-assisted image analysis tools is becoming a mainstream for automatic quantitative or semi-quantitative analysis for pathologists, including the discovery of novel predictive biomarkers [12]. However, a lot of challenges remain to be addressed for machine learning methods because to the high variability of quality of tissue preparation and digital acquisition, and in tissue phenotype. A central objective in digital pathology

is the accurate identification of cell or tissue of interest. For instance, computational staining of tumor tissue could be used for slide screening in order to increase the efficiency of pathologists. Automatically computed tumor maps could identify regions of interest for whole slide image classification [5], or be combined with automatic detection of lymphocytes [15] to further characterize the tumor and immune micro-environment for predicting treatment response [3].

Traditionally, image segmentation is tackled by leveraging pixel-wise or patch-wise ground-truth annotations [9]. This is highly problematic in digital pathology due to the colossal size of WSIs with respect to the biological components, implying that the annotation process is considerably time-consuming. Moreover, the high variance of clinical samples contributes on the deficiency of generalization, as illustrated in [4] where the front-runner solution of the CAMELYON16 challenge [2] has reportedly 4 times higher classification errors on the same task for in-house data from the same location.

A standard multiple instance learning (MIL) scheme [8] deals with classifying an entity (bag) from its constituents (instances). The MIL paradigm is particularly suited to histopathological image analysis due to its ability on reasoning on subsets of data (patches) that is often a computational necessity in histopathology. The general approach of MIL consists in learning a model that embeds each instance of a bag into a latent space. Then, the collection (usually of fixed size) of instances latent vectors is forwarded into an aggregating function which outputs the predicted bag probability, using different principles such as max-pooling [8], support vector machine [1], or even attention-based neural networks [11]. Recent large-scale histopathological studies provides promising classification solutions based on the MIL scheme [4–6]. Such approaches generally indicate whether a slide is non-neoplastic (normal), or the predicted subtype of apparent tumor tissue without accurately identifying the tumoral regions in the slide.

There are two ways to interpret multiple instance learning: MIL for classifying bags (or slides), or MIL for training an instance classifier model, apparent to bag segmentation. In particular, studies such as [4–6] use max-pooling MIL and its relaxed formulation [18] to first train an instance model, and then investigate various ways to combine instance predictions into a slide prediction. These works demonstrate that MIL schemes provide powerful formulations for the WSI classification, by reaching AUC for tumor versus normal slide classification higher than 0.99. However, these studies lack extensive evaluation for a more detailed MIL-driven segmentation performance since slide-based classification measures could lead to erroneous assessment regarding instance-level performance.

In this paper, we propose a weakly supervised segmentation scheme that is able to generate tumor segmentation models using annotations from the conventional clinical practice of pathologists' assessment. The contributions of this paper are: (i) a generic meta-algorithm, generating labels from WSI binary values intended to train detailed WSI segmentation models, (ii) a training scheme providing instance-level predictions, trained only with binary WSI annotations, (iii) the release of 6481 automatically generated tumor maps for publicly

available slides of The Cancer Genome Atlas (TCGA), an order of magnitude above previous released WSI tumor maps.

2   Weakly Supervised Learning for Tissue-Type Segmentation in Histopathological Images

Contextually, we consider a set $S = \{S_i\}$ of training whole slide images, where each slide S_i is associated with a label $T_i = \{0, 1\}$ where 1 refers to tumor and 0 to normal. More precisely, $T_i = 0$ indicates that there is no apparent tissue in the slide, and $T_i = 1$ indicates that some tissue is tumorous. The goal is to learn a tumor segmentation model, or a patch classifier, using only those binary annotations. To train a model in a fully supervised setup, a batch of patches $\{p_s\}$ is randomly sampled from a WSI along with their annotations computed beforehand. However, such microscopic annotations are impractical to obtain when the number of whole slides images is in the hundreds or thousands, which is necessary for good generalization. To deal with this limitation, the aim of the proposed framework is to generate a set of proxy patch-level ground truth labels by exploiting properties from the available global T_i labels.

By construction, a WSI with $T_i = 0$ indicates that all extracted patches are of normal class (0). In that case, a proxy-label filled with 0 provides perfect instance annotations, equivalent to a fully supervised learning scheme. However, in slides with $T_i = 1$, tumor tissue could possibly be in any extent on the S_i. Alternatively, a WSI with $T_i = 1$, normal tissue can theoretically cover no pixel up to the entire region in the slide except one pixel. We integrate this property by proposing a training scheme in which two parameters α and β are used for each training slide of $T_i = 1$ in the following deterministic process:

– assign a label 1 to the patches ensuring at least $\alpha\%$ are of class 1
– assign a label 0 to the patches ensuring at least $\beta\%$ are of class 0
– discard other patches from the computation of the loss signal

In such a setup, $\alpha\%$ represents the minimum assumed relative area of tumor tissue in the WSI, and similarly for $\beta\%$ with the normal tissue extent. Because the explicit process is deterministic, the framework is identified by the values of α and β. Noteworthy, (α, β) such that $(\alpha + \beta) > 100\%$ would produce contradictory proxy labels for $100 - (\alpha + \beta)\% > 0$ instances, which could impede training by diminishing the signal-to-noise ratio. Therefore, the possible space of these two parameters can be defined as $\mathbb{F} = \{(\alpha, \beta); \alpha > 0, \beta \geq 0, \alpha + \beta \leq 1\}$.

Formally, given a loss function L (e.g. binary cross-entropy), the formulated framework aims at minimizing the following empirical risk on S:

$$c_0 \cdot \sum_{\substack{S_i \in S; \\ T_i = 0}} \left[\underbrace{\sum_{p_s \sim S_i} L(f(p_s), 0)}_{R_{\mathrm{FP}}} \right]$$

$$+ c_1 \cdot \sum_{\substack{S_i \in S; \\ T_i = 1}} \left[\underbrace{\sum_{\substack{p_s \sim S_i; \\ p_s \in P(f(p_s); \alpha, 100)}} L(f(p_s), 1)}_{R_\alpha} + \underbrace{\sum_{\substack{p_s \sim S_i; \\ p_s \in P(f(p_s); 0, \beta)}} L(f(p_s), 0)}_{R_\beta} \right]$$

where $P(f(p_s); p_{\min}, p_{\max})$ is defined as the subset of patches $f(p_s)$ for which the predicted probability lies within the $p_{\min}th$ and the $p_{\max}th$ percentiles of the predictions $f(p_s)$, c_0 and c_1 are constants for batch averaging and class imbalance for both classes, and T_i refers to the binary ground-truth of S_i. Minimizing this empirical risk will guide models into recalling enough positive tumoral patches (R_α) per slide but not too much (R_β) while maintaining a low level of false positive in negative slides R_{FP}. The formulated approach is generic, in the sense that it can be used to train a large scope of machine learning models, including neural networks, with patch-based or pixel-wise predictions, and it can be coupled with most usual loss functions. It produces trained segmentation models, readily available to produce heatmaps without intervention of the formulated pipeline nor $\alpha\%$ and $\beta\%$ parameters.

3   Implementation Details and Dataset

3.1   Framework Setup and Architecture Details

We perform a benchmark of a representative population of the framework parameters space $\mathbb{F}$. Specifically, $\mathbb{F}$ is sampled starting from 0 with increment of 0.2 (or 20%) for both α and β (*e.g.* $(0,0)$, $(0, 0.2)$, $(0, 0.4)$ and so on), resulting in $\frac{6 \cdot 7}{2} = 21$ configurations. Of those, the 6 configurations with $\alpha = 0$ are discarded, as this would imply that the framework provides only 0 labels contradicting with the $T_i = 1$ assumption of our empirical risk formulation. At the end only 15 sampled configurations had been used.

Each configuration is used to train a ResNet50 architecture [10], which has been extensively used for histopathology image analysis in a multitude of tasks [14], and can be used without the global average pooling layer to yield 13×13 outputs per 224 pixel-wide input image. Pre-training is used with initialization on a well-performing snapshot on ImageNet [7]. At each epoch, each training slide is sampled once. Upon sampling, a batch size of 150 patches of size 224×224 are randomly sampled at 20x magnification in the tissue region of a WSI. Data augmentation is used independently on each image, with random rotations and flips while also applying channel-wise standard scaling from training averages and variances, and color jitter. The model is then concurrently inferred on the 150 patches of the batch, and a proxy-vector is constructed with the formulated pipeline as illustrated for 10 patches in Fig. 1. Specifically, the
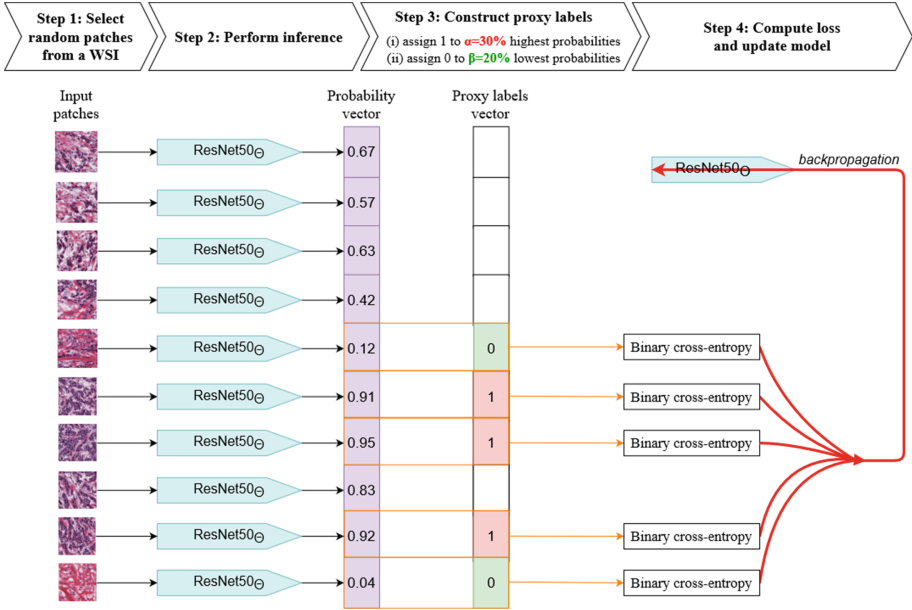
Fig. 1. Illustration of the processing of a batch of 10 patches from a positive WSI. A unique ResNet50 model with parameters θ is inferred on all images of the batch. For a given configuration of (α, β) (here, $(0.3, 0.2)$), these predictions are first used to create a proxy image-wise labels vector, then combined with the proxy label to compute batch error for further parameters θ update with backpropagation.

$\lfloor 150 \times \alpha \rfloor$ patches of highest probabilities are associated with a proxy value of 1, and the $\lfloor 150 \times \beta \rfloor$ patches of lowest probabilities with a proxy value of 0. A masking vector of size 150 is concurrently filled such that patches with no attributed label are discarded. The proxy vector is then coupled with the model's 150 predictions minus the discarded ones, in order to compute patch-wise binary cross-entropy loss which is then averaged and retro-propagated across all the non masked predictions. The error signal is used to tune model parameters using Adam optimizer with learning rate of 10^{-4} and default momentum parameters. c_0 and c_1 are set to 1. Each configuration is trained for 20 epochs on 2 V100 NVIDIA GPU, for a training time of $\sim$16 h, or a total benchmark training time of $\sim$240 h. The code is implemented with pytorch 1.5.0 and torchvision 0.4.0 on python3.6.

3.2   Dataset

The dataset consists of 6481 flash-frozen whole slide images from TCGA, issued from kidney (2334), bronchus and lung (2168) and breast (1979) WSIs locations. These locations were selected on TCGA as the first 3 indexed, while no slide filtering nor slide selection has been performed to be coherent with standard

clinical practices. This dataset was divided in training, validation and testing sets on a case basis, with 65%, 15%, and 20% of cases respectively. For the rest of the paper this testing set is denoted as "In-distribution". Each selected configuration is trained using the training set, with hyper-parameters optimized on the validation set. Then, their performance is assessed on the testing set. For extensive quantitative performance assessment, expert pathologists annotated 130 slides from this testing set (45 breast, 40 kidney, 45 bronchus and lung), thus measuring in-distribution generalization. Annotations were computed at 20x magnification by a junior pathologist on a in-house annotation tool by contouring tumor tissue regions which were then filled, and were modified until validation by a senior pathologist.

The same protocol was applied on additional slides extracted from locations which are not used in the previous cohort. Specifically, 35 WSIs from colon, 35 from ovary and 30 from corpus uteri are pixel-wise annotated and used to measure generalization performance of models to unseen tissue environments, which we denote as the "Out-of-location" testing set. We pinpoint that these annotations were not used during training nor validation, but only to assess testing segmentation performance of the produced models. For training, we use diagnostic labels extracted directly from TCGA, for which each slide name contain a code indicating the type of specimen (e.g. "Solid Tissue Normal" or "Primary Solid Tumor")[1]. Notably, normal slides are explicitly discerned from slide with apparent pathological tissue. In such context, each slide is associated with a binary value indicating whether tumor tissue is apparent in the slide, or whether the slide is of normal tissue only. To further compare with results from the community, we infer all models on the PatchCamelyon dataset [17]. The dataset consists in 96×96 patches extracted from formalin-fixed, paraffin-embedded (FFPE) tissues from sentinel lymph nodes at 10x magnification. In PatchCamelyon, images are labeled as 1 if their 32×32 center region contains at least 1 pixel of tumor cell, otherwise 0. To accommodate with the 224×224 input at 20x magnification of the learned models, these images were bi-linearly upsampled twice and padded with 0. This testing set is particularly challenging for the benchmarked models, since they are not trained on FFPE slides, which are visually highly different from flash-frozen ones. Besides, the trained models did not include any tissue extracted from sentinel lymph nodes, highlighting the generalization challenge of the proposed framework.

4   Results and Discussion

For performance assessment, all 15 trained ResNet50 models are inferred on the testing slides. The resulting heatmaps are compared to segmentations maps provided by the pathologists. All configurations are found to converge to suprandom In-distribution performance, except for the two extreme configurations ($\alpha = 1, \beta = 0$) and $(0.2, 0.8)$, as displayed in Table 1. The average In-distribution AUC is 0.675 ± 0.132, with optimal AUC of 0.804 for ($\alpha = 0.2, \beta = 0.2$).

---

[1] https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/sample-type-codes.

Precision and recall are extracted after threshold selection on validation set and displayed in Fig. 2. The α parameter seems to influence the recall at the expense of precision. Upon performance introspection by location, all configurations report the worse performance for the location bronchus and lung, with twice as much AUC error compared to kidney and breast locations. Concerning the Out-of-distribution cohort, the average AUC is 0.679 ± 0.154, which is close to In-distribution performance, although lower when omitting bronchus and lung location from the latter. There is no evident pattern for configurations that yield improved Out-of-location results.

Table 1. Pixelwise AUC for the 15 (α, β) framework configurations on the hold-out testing set (In-distribution) and the testing set from locations unseen in training (Out-of-location). Grey results take background into account, black ones are computed by completely discarding background from performance computation.

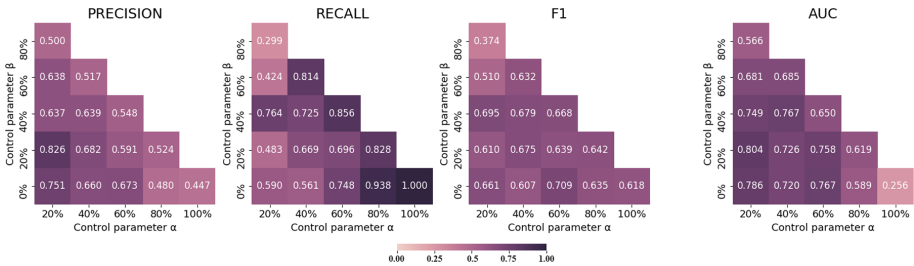|	$\alpha =$ 0.2					0.4				0.6			0.8		1.0
$\beta =$	0	0.2	0.4	0.6	0.8	0	0.2	0.4	0.6	0	0.2	0.4	0	0.2	0
In-distribution	.786	.804	.749	.681	.566	.720	.726	.767	.685	.766	.758	.650	.589	.619	.256
	.964	.952	.960	.930	.874	.967	.953	.959	.935	.957	.960	.940	.946	.946	.926
Out-of-location	.866	.732	.709	.583	.257	.783	.710	.762	.658	.790	.787	.673	.785	.695	.404
	.984	.974	.972	.958	.917	.980	.972	.978	.966	.981	.980	.968	.980	.970	.933

Fig. 2. Quantitative results for the 15 benchmarked configurations on the hold-out testing set (In-distribution) from bronchus and lung, kidney, and breast locations. Each subplot (4 in total) displays a pixelwise measure, as indicated in its sup-title, for each configuration in a matrix format. AUC: area under the ROC curve.

Some visual representations of two different samples testing are presented on Fig. 3. In particular, in the figure we present the WSI image together with the pixel-wise annotations of the pathologist. Moreover, different segmentation maps depending on the configuration are also presented. It can observed that there are 3 or 4 configurations that are close to the expert's annotations. These configurations are in line with our quantitative results. Additional post-processing strategies would potentially boost the performance of our framework.
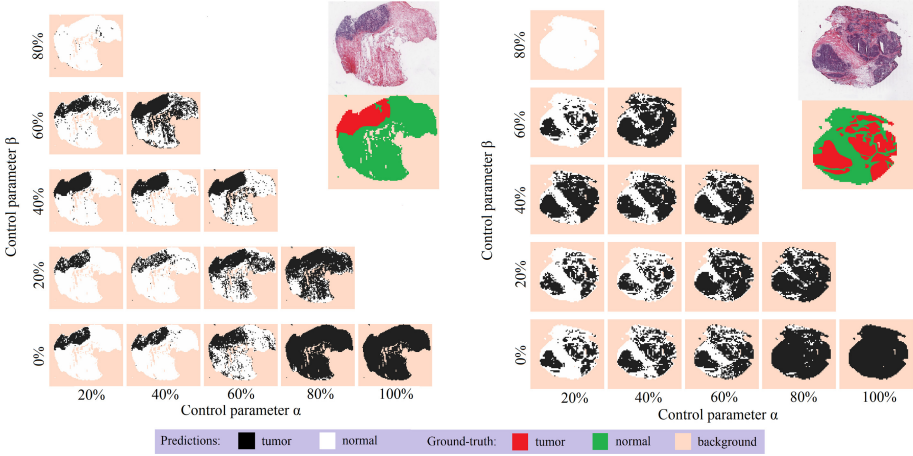
Fig. 3. Unfiltered predicted tumor maps on hold-out testing samples for the 15 benchmarked framework configurations. 2 WSI and their corresponding results are displayed in a matrix-format. The red and green images are the ground-truth.

To test the generalization of our method, we performed also experiments on the PatchCamelyon dataset [17]. In particular we found the most of the configurations (12 out of the 15) reporting quite low AUC, between 0.428 and 0.612. However, 3 configurations are found to generalize to some extent, that is $(\alpha = 0.2, \beta = 0)$ with 0.672 AUC, $(0.2, 0.2)$ with 0.802, and $(0.4, 0)$ with 0.758. Although these results are far from report AUC of 0.963 obtained with fully supervised models specifically trained on this dataset [17], the results suggest the presented framework could provide models which can grasp generic discriminative cancer features from multiple types of slides in broad biological context.

5   Conclusion and Future Works

In this paper we propose a weakly supervised model which provides segmentation maps of WSIs, trained only with binary annotations over the entire WSIs. From our experiments we saw that usually 3 to 4 configurations are expected to yield respectively high precision, high recall, and high overall performance for WSIs of different organs and tumor coverage. The findings in this paper highlight the potential of weakly supervised learning in histopathological image segmentation, which is known to be heavily impeded by the annotation bottleneck. With the complete open-source releases of both the complete WSI pre-processing pipeline, the presented training framework, as well as the inference pipeline, the presented approach can be used off-the-shelf for pan-cancer tumor segmentation using the entire 18k flash-frozen WSI of TCGA, or other type of tissue segmentation such as necrosis or stromal tissue. The public release of 6481 automatically generated tumor maps, with an expected AUC above 0.932, should lower the barrier of

entry to pathomics by bypassing tumor annotation efforts. All code and results can be found at https://github.com/marvinler/tcga_segmentation.

There are many ways to fine-tune a segmentation model using the formulated framework, such as with more appropriate deep learning architectures or with more extensive hyper-parameters optimization. We believe the most impactful future works will revolve around the proxy-generation labels from more sophisticated slide labels which would yield higher information while remaining fast to obtain, essentially trading annotation time for performance.

References

1. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: Advances in Neural Information Processing Systems, pp. 577–584 (2003)
2. Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., et al.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama **318**(22), 2199–2210 (2017)
3. Binnewies, M., Roberts, E.W., Kersten, K., Chan, V., Fearon, D.F., Merad, M., Coussens, L.M., Gabrilovich, D.I., Ostrand-Rosenberg, S., Hedrick, C.C., et al.: Understanding the tumor immune microenvironment (time) for effective therapy. Nat. Med. **24**(5), 541–550 (2018)
4. 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. Nat. Med. **25**(8), 1301–1309 (2019)
5. Campanella, G., Silva, V.W.K., Fuchs, T.J.: Terabyte-scale deep multiple instance learning for classification and localization in pathology (2018). arXiv preprint arXiv:1805.06983
6. Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., Tsirigos, A.: Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat. Med. **24**(10), 1559 (2018)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
8. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artif. Intell. **89**(1–2), 31–71 (1997)
9. Guo, Y., Liu, Y., Georgiou, T., Lew, M.S.: A review of semantic segmentation using deep neural networks. Int. J. Multimedia Inf. Retrieval **7**(2), 87–93 (2017). https://doi.org/10.1007/s13735-017-0141-z
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
11. Ilse, M., Tomczak, J.M., Welling, M.: Attention-based deep multiple instance learning (2018). arXiv preprint arXiv:1802.04712
12. Janowczyk, A., Madabhushi, A.: Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. J. Pathol. Inform. **7**, 29 (2016)

13. Jara-Lazaro, A.R., Thamboo, T.P., Teh, M., Tan, P.H.: Digital pathology: Exploring its applications in diagnostic surgical pathology practice. Pathology **42**(6), 512–518 (2010)
14. Mormont, R., Geurts, P., Marée, R.: Comparison of deep transfer learning strategies for digital pathology. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 2262–2271 (2018)
15. Saltz, J., et al.: Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. Cell Reports **23**(1), 181–193 (2018)
16. Stathonikos, N., Veta, M., Huisman, A., van Diest, P.J.: Going fully digital: Perspective of a dutch academic pathology lab. J. Pathol. Inform. **4**, 15 (2013)
17. Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., Welling, M.: Rotation equivariant CNNs for digital pathology. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 210–218. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_24
18. Zhu, W., Lou, Q., Vang, Y.S., Xie, X.: Deep multi-instance networks with sparse label assignment for whole mammogram classification. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 603–611. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_69