Classification and Disease Localization in Histopathology Using Only Global Labels: A Weakly-Supervised Approach

Pierre Courtiol,[*] Eric W. Tramel,[*] Marc Sanselme, and Gilles Wainrib
Owkin, Inc.
New York City, NY
{firstname.lastname}@owkin.com

Abstract

Analysis of histopathology slides is a critical step for many diagnoses, and in particular in oncology where it defines the gold standard. In the case of digital histopathological analysis, highly trained pathologists must review vast whole-slide-images of extreme digital resolution ($100,000^2$ pixels) across multiple zoom levels in order to locate abnormal regions of cells, or in some cases single cells, out of millions. The application of deep learning to this problem is hampered not only by small sample sizes, as typical datasets contain only a few hundred samples, but also by the generation of ground-truth localized annotations for training interpretable classification and segmentation models. We propose a method for disease localization in the context of weakly supervised learning, where only image-level labels are available during training. Even without pixel-level annotations, we are able to demonstrate performance comparable with models trained with strong annotations on the Camelyon-16 lymph node metastases detection challenge. We accomplish this through the use of pre-trained deep convolutional networks, feature embedding, as well as learning via top instances and negative evidence, a multiple instance learning technique from the field of semantic segmentation.

1 Introduction

Histopathological image analysis (HIA) is a critical element of diagnosis in many areas of medicine, and especially in oncology, where it defines the gold standard metric. Recent works have sought to leverage developments in machine learning (ML) to aid pathologists in disease detection tasks, but the majority of these techniques require localized annotation masks as training data. These annotations are even more costly to obtain than the original diagnosis, as pathologists must spend time to assemble pixel-by-pixel segmentation maps of diseased tissue at extreme resolution, thus HIA datasets with annotations are very limited in size. Additionally, such localized annotations may not be available when facing new problems in HIA, such as new disease subtybe classification, prognosis estimation, or drug response prediction. Thus, the critical question for HIA is: can one design a learning architecture which achieves accurate classification with no additional localized annotation? A successful technique would be able to train algorithms to assist pathologists during analysis and could also be used to identify previously unknown structures and regions of interest.

Indeed, while histopathology is the gold standard diagnostic in oncology, it is extremely costly, requiring many hours of focus from pathologists to make a single diagnosis (Litjens et al., 2016; Weaver, 2010). Additionally, as correct diagnosis for certain diseases requires pathologists to identify a few cells out of millions, these tasks are akin to “finding a needle in a haystack.” In digital

[*] Equal contribution to this work.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

pathology with whole-slide-imaging (WSI) (Yagi & Gilbertson, 2005; Snead et al., 2016), highly trained and skilled pathologists review digitally captured microscopy images from prepared and stained tissue samples in order to make diagnoses. Digital WSI are massive datasets, consisting of images captured at multiple zoom levels. At the greatest magnification levels, a WSI may have a digital resolution upwards of 100 thousand pixels in both dimensions. However, since localized annotations are very difficult to obtain, one is often left with image-level global labels. Thus, obtaining predictions localized within the WSI becomes a difficult task commonly referred to as *weakly-supervised learning*.

In this paper, we propose CHOWDER[2], an approach for the interpretable prediction of general localized diseases in WSI with only weak, whole-image disease labels and without any additional expert-produced localized annotations, i.e. per-pixel segmentation maps, of diseased areas within the WSI. We accomplish this through the use of a novel aggregation technique performed on features extracted by a pre-trained DCNN on the tile-level. Notably, while the approach we propose makes use of a pre-trained model, the entire procedure is a true end-to-end classification technique. Thus, pre-trained DCNN feature-extraction layers can be fine-tuned to the context of haematoxylin and eosin (H&E) stained WSI. We demonstrate, using only whole-slide labels, performance comparable to top-10 ranked methods trained with strong, pixel-level labels on the Camelyon-16 challenge dataset, while also producing disease segmentation that closely matches ground-truth annotations. We also present results for diagnosis prediction on WSI obtained from The Cancer Genome Atlas (TCGA), where strong annotations are not available and diseases may not be strongly localized in the tissue sample.

2 Baseline: Global Feature Aggregation

Given their scale, it is necessary to process WSI in a piecemeal fashion, taking in image data for a given magnification level tile-by-tile. Following the pre-processing pipeline described in Appendix A, extracting tile-level features produces a bag of feature vectors which one attempts to use for classification against the known image-wide label. The dimension of these local descriptors is $M^{\mathrm{S}} \times P$, where P is the number of features output from the pre-trained image DCNN and M^{S} is the number of sampled tiles. Approaches such as Bag-of-visual-words (BoVW) or VLAD (Jégou et al., 2010) could be chosen as a baseline aggregation method to generate a single image-wide descriptor of size $P \times 1$, but would require a huge computational power given the dimensionality of the input. Instead, we will try two common approaches for the aggregation of local features, specifically, the `MaxPool` and `MeanPool`.

After applying these pooling methods over the axis of tile indices, one obtains a single feature descriptor for the whole image. Other pooling approaches have been used in the context of HIA, including Fisher vector encodings (Song et al., 2017) and $p-$norm pooling (Xu et al., 2017). However, as the reported effect of these aggregations is quite small, we don't consider these approaches when constructing our baseline approach. After aggregation, a classifier can be trained to produce the desired diagnosis labels given the global WSI aggregated descriptor. For our baseline method, we use a logistic regression for this final prediction layer of the model.

3 CHOWDER Method

In experimentation, we observe that the baseline approach of the previous section works well for *diffuse* disease, which is evidenced in the results of Table 1 for `TCGA-Lung`. Here, *diffuse* implies that the number of disease-containing tiles, pertinent to the diagnosis label, are roughly proportional to the number of tiles containing healthy tissue. However, if one applies the same approach to different WSI datasets, such as `Camelyon-16`, the performance significantly degrades. In the case of `Camelyon-16`, the diseased regions of most of the slides are highly localized, restricted to a very small area within the WSI. When presented with such imbalanced bags, simple aggregation approaches for global slide descriptors will overwhelm the features of the disease-containing tiles. Instead, we propose an improvement of the WELDON method (Durand et al., 2016), adapting it for use in HIA. As in Durand et al. (2016), rather than creating a global slide descriptor by aggregating

[2] Classification of HistOpathology with Weak supervision via Deep fEature aggRegation

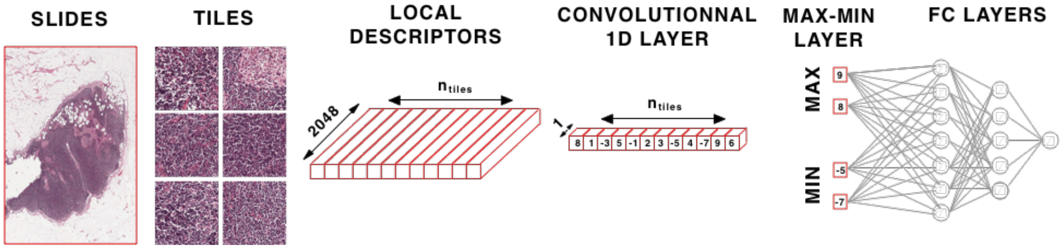

Figure 1: Description of the CHOWDER architecture (for $R = 2$) for WSI classification via MLP on operating on top positive and negative instances shown for a single sample mini-batch sample.

all tile features, instead a *multiple instance learning* (MIL) approach is used that combines both top-instances as well as negative evidence. A visual description of approach is given in Fig. 1.

First, a set of one-dimensional embeddings for the $P = 2048$ ResNet-50 features are calculated via J one-dimensional convolutional layers strided across the tile index axis. For tile t with features $\mathbf{k}_t$, the embedding according to kernel j is calculated as $e_{j,t} = \langle \mathbf{w}_j, \mathbf{k}_t \rangle$. Notably, the kernels $\mathbf{w}_j$ have dimensionality P. This one-dimensional convolution is, in essence, a shortcut for enforcing a fully-connected layer with tied weights across tiles, i.e. the same embedding for every tile (Durand et al., 2016). In our experiments, we found that the use of a single embedding, $J = 1$, is an appropriate choice for WSI datasets when the number of available slides is small (< 1000). In this case, choosing $J > 1$ will decrease training error, but will *increase* generalization error. Avoiding overtraining and ensuring model generality remains a major challenge for the application of WSL to WSI datasets.

After feature embedding, we now have a $M^{\mathrm{S}} \times 1$ vector of local tile-level (*instance*) descriptors. As in Durand et al. (2016), these instance descriptors are sorted by value. Of these sorted embedding values, only the top and bottom R entries are retained, resulting in a vector of $2R \times 1$ entries to use for diagnosis classification. This can be easily accomplished through a `MinMax` layer on the output of the one-dimensional convolution layer. The purpose of this layer is to take not only the top instances region but also the negative evidence, i.e. the region which best support the absence of the class. During training, the back-propagation runs only through the $2R$ selected tiles, the positive and negative evidence. When applied to WSI, the `MinMax` serves as a powerful tile selection procedure.

In the WELDON architecture, the last layer consists of a sum applied over the $2R \times 1$ output from the `MinMax` layer. However, we find that this approach can be improved for WSI classification. We investigate the possibility of richer interactions between the top and bottom instances by instead using an MLP as the final classifier. In our implementation of CHOWDER, we use an MLP with two fully connected layers of 200 and 100 neurons with sigmoid activations.

Lastly, while the CHOWDER architecture is trained predict WSI classification, the trained model may be modified during inference to permit tumor localization, as well. Specifically, for a set of tiles drawn from a given WSI, the output of the feature embedding layer is used as an indicator of tile-level classification. In the specific case of $J = 1$, the magnitude of the scalar value output for each tile is used to represent the relative strength of classification of the tile-level instance between binary classes (e.g. “healthy”, “tumor”). For high-quality localization maps, a large number of overlapping tiles may be drawn and their inferred instance-level values fused together into a single map.

4 Experimental Results

For pre-processing, we fix a single tile scale for all methods and datasets. We chose a fixed zoom level of 0.5 μm/pixel, which corresponds to $\ell = 0$ for slides scanned at 20x magnification, or $\ell = 1$ slides scanned at 40x magnification. Next, since WSI datasets often only contain a few hundred images, far from the millions images of ImageNet dataset, strong regularization required prevent over-fitting. We applied ℓ_2-regularization of 0.5 on the convolutional feature embedding layer and dropout on the MLP with a rate of 0.5. However, these values may not be the global optimal, as we did not apply any hyper-parameter optimization to tune these values. To optimize the model parameters, we use Adam (Kingma & Ba, 2014) to minimize the binary cross-entropy loss over 30 epochs with a mini-batch size of 10 and with learning rate of 0.001.

To reduce variance and prevent over-fitting, we trained an ensemble of E CHOWDER networks which only differ by their initial weights. The average of the predictions made by these E networks establishes the final prediction. Although we set $E = 10$ for the results presented in Table 1, we used a larger ensemble of $E = 50$ with $R = 5$ to obtain the best possible model and compare our method to those presented in Table 1. We also use an ensemble of $E = 10$ when reporting the results for WELDON. As the training of one epoch requires about 30 seconds on our available hardware, the total training time for the ensemble took just over twelve hours. While the ResNet-50 features were extracted using a GPU for efficient feed-forward calculations, the CHOWDER network is trained on CPU in order to take advantage of larger system RAM sizes, compared to on-board GPU RAM. This allows us to store all the training tiles in memory to provide faster training compared to a GPU due to reduced transfer overhead.

TCGA. The public Cancer Genome Atlas (TCGA) provides approximately 11,000 tissue slides images of cancers of various organs[3]. For our first experiment, we selected 707 lung cancer WSIs (`TCGA-Lung`), which were downloaded in March 2017. Subsequently, a set of new lung slides have been added to TCGA, increasing the count of lung slides to 1,009. Along with the slides themselves, TCGA also provides labels representing the type of cancer present in each WSI. However, no local segmentation annotations of cancerous tissue regions are provided. The pre-processing step extracts 1,411,043 tiles and their corresponding representations from ResNet-50. The task of these experiments is then to predict which type of cancer is contained in each WSI: adenocarcinoma or squamous cell carcinoma. We evaluate the quality of the classification according to the area under the curve (AUC) of the receiver operating characteristic (ROC) curve generated using the raw output predictions.

As expected in the case of diffuse disease, the advantage provided by CHOWDER is slight as compared to the `MeanPool` baseline, as evidenced in Table 1. Additionally, as the full aggregation techniques work quite well in this setting, the value of R does not seem to have a strong effect on the performance of CHOWDER as it increases to $R = 100$. In this setting of highly homogenous tissue content, we can expect that global aggregate descriptors are able to effectively separate the two classes of carcinoma.

Camelyon-16. For our second experiment, we use the `Camelyon-16` challenge dataset[4], which consists of 400 WSIs taken from sentinel lymph nodes, which are either healthy or exhibit metastases of some form. In addition to the the WSIs themselves, as well as their labeling (`healthy`, `contains-metastases`), a segmentation mask is provided for each WSI which represents an expert analysis on the location of metastases within the WSI. Human labeling of sentinel lymph node slides is known to be quite tedious, as noted in Litjens et al. (2016); Weaver (2010). Teams participating in the challenge had access to, and utilized, the ground-truth masks when training their diagnosis prediction and tumor localization models. For our approach, we set aside the masks of metastasis locations and utilize only diagnosis labels. Furthermore, many participating teams developed a post-processing step, extracting handcrafted features from predicted metastasis maps to improve their segmentation. No post-processing is performed for the presented CHOWDER results, the score is computed directly from the raw output of the CHOWDER model. We also conduct a set of experiments on `Camelyon-16` using random train-test cross-validation (CV) splits, respecting the same training set size as in the original competition split.

The `Camelyon-16` dataset is evaluated on two different axes. First, the accuracy of the predicted label for each WSI in the test set is evaluated according to AUC. Second, the accuracy of metastasis localization is evaluated by comparing model outputs to the ground-truth expert annotations of metastasis location. This segmentation accuracy is measured according to the free ROC metric (FROC), which is the curve of metastasis detection sensitivity to the average number of also positives. As in the Camelyon challenge, we evaluate the FROC metric as the average detection sensitivity at the average false positive rates 0.25, 0.5, 1, 2, 4, and 8.

In Table 1, we see the classification performance of our proposed CHOWDER method, for $E = 10$, as compared to both the baseline aggregation techniques, as well as the WELDON approach. In the case of WELDON, the final MLP is not used and instead a summing is applied to the `MinMax` layer. The value of R retains the same meaning in both cases: the number of both high and low

[3] `https://portal.gdc.cancer.gov/legacy-archive`

[4] `https://camelyon16.grand-challenge.org`

	Camelyon		
Method	*CV*	*Competition*	**TCGA**
BASELINE			
`MaxPool`	0.749	0.655	0.860
`MeanPool`	0.802	0.530	0.903
WELDON			
$R = 1$	0.782	0.765	—
$R = 10$	0.832	0.670	—
$R = 100$	0.809	0.600	—
$R = 300$	0.761	0.573	—
CHOWDER			
$R = 1$	0.809	0.821	0.900
$R = 5$	**0.903**	**0.858**	—
$R = 10$	0.900	0.843	**0.915**
$R = 100$	0.870	0.775	0.909
$R = 300$	0.837	0.652	—

Rank	**Team**	**AUC**
1	HMS & MIT	0.9935
2	HMS-MGH	0.9763
3	HMS-MGH	0.9650
	. . .	
10	DeepCare Inc.	0.8833
	CHOWDER	**0.8706**
11	Indep. DE	0.8654
	. . .	

Rank	**Team**	**FROC**
1	HMS & MIT	0.8074
2	HMS-MGH	0.7600
3	HMS-MGH	0.7289
	. . .	
17	SIT	0.3385
	CHOWDER	**0.3103**
18	Warwick-QU	0.3052
	. . .	

Table 1: *Left:* Classification (AUC) results for the `Camelyon-16` and `TCGA-Lung` datasets for CHOWDER, WELDON, and the baseline approach. For `Camelyon-16`, we present two scores, one for the fixed competition test split of 130 WSIs, and one for a cross-validated average over 3 folds (*CV*) on the 270 training WSIs. For `TCGA-Lung`, we present scores as a cross-validated average over 5 folds. *Right:* Final leader boards for `Camelyon-16` competition. All competition methods had access to the full set of strong annotations for training their models. In contrast, our proposed approach only utilizes image-wide diagnosis levels and obtains comparable performance as top-10 methods.

scoring tiles to pass on to the classification layers. We test a range of values R for both WELDON and CHOWDER. We find that over all values of R, CHOWDER provides a significant advantage over both the baseline aggregation techniques as well as WELDON. We also note that the optimal performance can be obtained without using a large number of discriminative tiles, i.e. $R = 5$.

For CHOWDER, we also propose tumor localization via the full set of outputs from the convolutional feature embedding layer. These are then sorted and thresholded according to value τ such that tiles with an embedded value larger than τ are classified as diseased and those with lower values are classified as healthy. We show an example of disease localization produced by CHOWDER in Appendix A. Here, we see that CHOWDER is able to very accurately localize the tumorous region in the WSI even though it has only been trained using global slide-wide labels and without any local annotations. While some potential false detections occur outside of the tumor region, we see that the strongest response occurs within the tumor region itself.

We also present in Table 1 our performance as compared to the public Camelyon leader boards for $E = 50$. We are able to obtain an effective 11^{th} and 18^{th} place rank for classification and localization, respectively, *but without using any of the ground-truth disease segmentation maps*. This is a remarkable result, as the winning approach of Wang et al. (2016) required tile-level disease labels derived from expert-provided annotations in order to train a full 27-layer GoogLeNet (Szegedy et al., 2015) architecture for tumor prediction.

5 Discussion

We have shown that using state-of-the-art techniques from MIL in computer vision, such as the top instance and negative evidence approach of (Durand et al., 2016), one can construct an effective technique for diagnosis prediction *and* disease location for WSI in histopathology without the need for expensive localized annotations produced by expert pathologists. By removing this requirement, we hope to accelerate the production of computer-assistance tools for pathologists to greatly improve the turn-around time in pathology labs and help surgeons and oncologists make rapid and effective patient care decisions. This also opens the way to tackle problems where expert pathologists may not know precisely where are the relevant zones in the images, for instance for prognosis estimation or prediction of drug response tasks. The ability of our approach to discover associated regions of interest without prior localized annotations hence appears as a novel discovery approach for the field of pathology. Moreover, using the suggested localization from CHOWDER, one may considerably speed up the process of obtaining ground-truth localized annotations.

References

Francesco Ciompi, Oscar Guessing, Babak Ehteshami Bejnordi, Gabriel Silva de Souza, Alexi Baidoshvili, Geert Litjens, Bram van Ginneken, Iris Nagtegaal, and Jeroen van der Laak. The importance of stain normalization in colorectal tissue classification with convolutional networks. arXiv Preprint [cs.CV]:1702.05931, 2017.

Thibault Durand, Nicolas Thome, and Matthieu Cord. WELDON: Weakly supervised learning of deep convolutional neural networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 4743–4752, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2016.

Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *Proc. IEEE Conf. on Computer Vision and Pattern Reconition*, 2010.

Adnan Mujahid Khan, Nasir Rajpoot, Darren Treanor, and Derek Magee. A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. *IEEE Trans. Biomedical Engineering*, 61(6), 2014.

Diederik P. Kingma and Jimmy Ba. ADAM: A method for stochastic optimization. arXiv Preprint [cs.LG]:1412.6980, 2014.

Geert Litjens, Clara I. Sanchez, Nadya Timofeeva, Meyke Hermsen, Iris Nagtegaal, Iringo Kovacs, Christina Hulsbergen van de Kaa, Peter Bult, Bram van Ginneken, and Jeroen van der Laak. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific Reports*, (6), 2016.

Dennis Nikitenko, Michael A. Wirth, and Kataline Trudel. Applicability of white-balancing algorithms to restoring faded colour slides: An empirical evaluation. *Journal of Multimedia*, 2008.

N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

David R J Snead, Yee-Wah Tsang, Aisha Meskiri, Peter K Kimani, Richard Crossman, Nasir M Rajpoot, Elaine Blessing, Klaus Chen, Kishore Gopalakrishnan, Paul Matthews, Navid Momtahan, Sarah Read-Jones, Shatrughan Sah, Emma Simmons, Bidisa Sinha, Sari Suortamo, Yen Yeo, Hesham El Daly, and Ian A Cree. Validation of digital pathology imaging for primary histopathological diagnosis. *Histopathology*, 68(7): 1063–1072, 2016.

Yang Song, Ju Jia Zou, Hang Chang, and Weidong Cai. Adapting fisher vectors for histopathology image classification. In *Proc. IEEE Int. Symp. on Biomedical Imaging*, 2017.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1–9, 2015.

Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H. Beck. Deep learning for identifying metastatic breast cancer. arXiv Preprint [q-bio.QM]:1606.05718, 2016.

D. L. Weaver. Pathology evaluation of sentinel lymph nodes in breast cancer: Protocol recommendations and rationale. *Mod. Pathol.*, 23(Suppl 2):S26–S32, 2010.

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

Yukako Yagi and John R. Gilbertson. Digital imaging in pathology: The case for standardization. *Journal of Telemedicine and Telecare*, 11(3):109–116, 2005.

A WSI Pre-Processing

Tissue Detection. It is possible that large regions of a WSI may contain no tissue at all. To extract only tiles with content relevant to the task, we use the same approach as Wang et al. (2016), namely, Otsu's method (Otsu, 1979) applied to the the hue and saturation channels of the image after transformation into the HSV color space to produce two masks which are then combined to produce the final tissue segmentation. Subsequently, only tiles within the foreground segmentation are extracted for training and inference.

Color Normalization. According to Ciompi et al. (2017), stain normalization is an important step in HIA since the result of the H&E staining procedure can vary greatly between any two slides. We utilize a simple histogram equalization algorithm consisting of left-shifting RGB channels and subsequently rescaling them to $[0, 255]$, as proposed in Nikitenko et al. (2008). In this work, we place a particular emphasis on the tile aggregation method rather than color normalization, so we did not make use of more advanced color normalization algorithms, such as Khan et al. (2014).

Tiling. The tiling step is necessary in histopathology analysis. Indeed, due to the large size of the WSI, it is computationally intractable to process the slide in its entirety. For example, on the highest resolution zoom level, denoted as *scale 0*, for a fixed grid of non-overlapping tiles, a WSI may possess more than 200,000 tiles of 224×224 pixels. Because of the computational burden associated with processing the set of all possible tiles, we instead turn to a uniform random sampling from the space of possible tiles. Additionally, due to the large scale nature of WSI datasets, the computational burden associated with sampling potentially overlapping tiles from arbitrary locations is a prohibitive cost for batch construction during training.

Instead, we propose that all tiles from the non-overlapping grid should be processed and stored to disk prior to training. As the tissue structure does not exhibit any strong periodicity, we find that sampling tiles along a fixed grid without overlapping provides a reasonably representative sampling while maximizing the total sampled area.

Given a target scale $\ell \in \{0, 1, \ldots, L\}$, we denote the number of possible tiles in WSI indexed by $i \in \{1, 2, \ldots, N\}$ as $M^{\mathrm{T}}_{i,\ell}$. The number of tiles sampled for training or inference is denoted by $M^{\mathrm{S}}_{i,\ell}$ and is chosen according to

$$M^{\mathrm{S}}_{i,\ell} = \min\left(M^{\mathrm{T}}_{i,\ell}\,,\ \max\left(M^{\mathrm{T}}_{\min},\ \frac{1}{2} \cdot \bar{M}^{\mathrm{T}}_{\ell}\right)\right), \tag{1}$$

where $\bar{M}^{\mathrm{T}}_{\ell} = \frac{1}{N}\sum_i M^{\mathrm{T}}_{i,\ell}$ is the empirical average of the number of tiles at scale ℓ over the entire set of training data.

Feature Extraction. We make use of the ResNet-50 (He et al., 2016) architecture trained on the ImageNet natural image dataset. In empirical comparisons between VGG or Inception architectures, we have found that the ResNet architecture provides features more well suited for HIA. Additionally, the ResNet architecture is provided at a variety of depths (ResNet-101, ResNet-152). However, we found that ResNet-50 provides the best balance between the computational burden of forward inference and richness of representation for HIA.

In our approach, for every tile we use the values of the ResNet-50 pre-output layer, a set of $P = 2048$ floating point values, as the feature vector for the tile. Since the fixed input resolution for ResNet-50 is 224×224 pixels, we set the resolution for the tiles extracted from the WSI to the same pixel resolution at every scale ℓ.

A Further Results

A.1 ROC Curves for Camelyon Competition Split

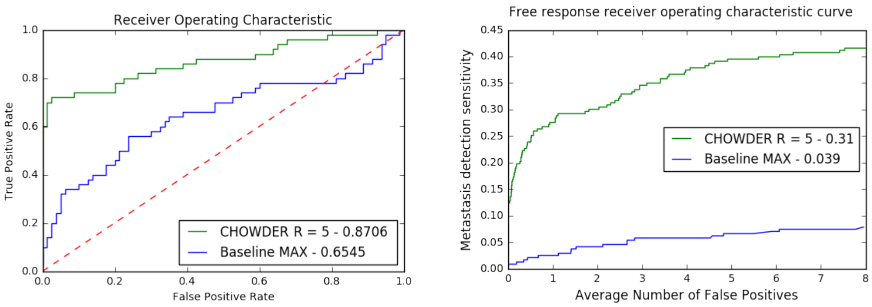

Figure 2: Performance curves for `Camelyon-16` dataset for both classification and segmentation tasks for the different tested approaches. *Left:* ROC curves for the classification task. *Right:* FROC curves for lesion detection task.

A.2 Tumor Localization Examples

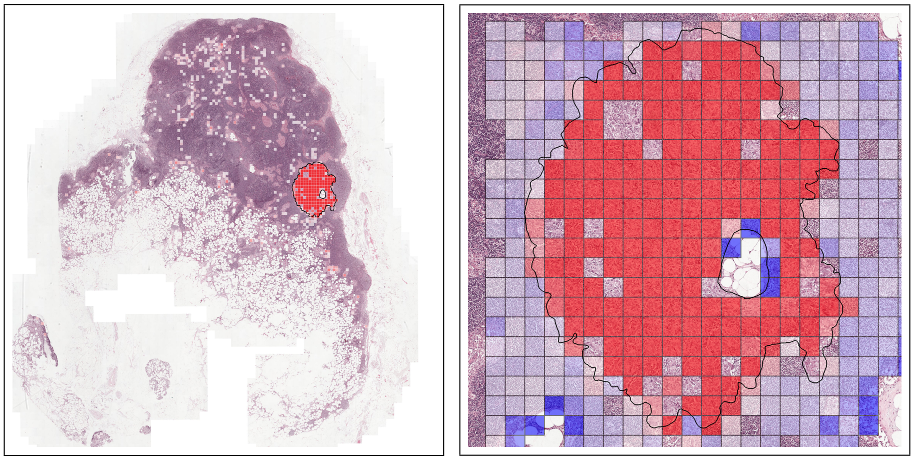

Figure 3: Visualization of metastasis detection on test image 27 of the `Camelyon-16` dataset using our proposed approach. **Left:** Full WSI at zoom level 6 with ground truth annotation of metastases shown via black border. Tiles with feature embeddings larger than 0 are classified as hits and color coded according to their magnitude, with red mapped to strong metastasis detection. **Right:** Detail of metastases at zoom level 2 overlaid with classification output of our proposed approach. Here, the output of all tested tiles are shown, with the feature embeddings mapped to a blue-white-red colormap, with blue and red mapped to strong healthy and metastasis detection, respectively. Tiles without color were not included when randomly selecting tiles for inference.

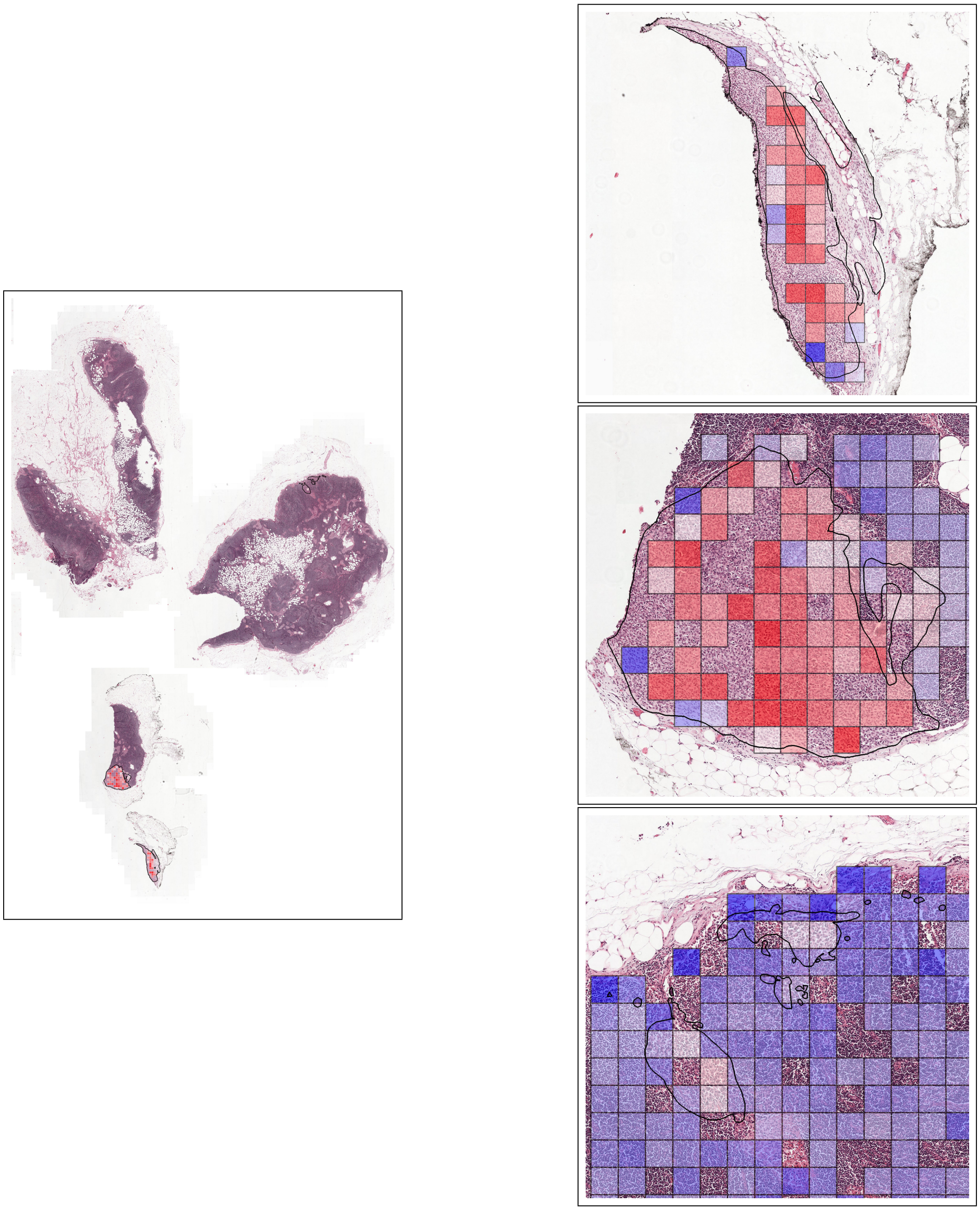

Figure 4: Visualization of metastasis detection on test image 2 of the `Camelyon-16` dataset using our proposed approach. **Left:** Full WSI at zoom level 6 with ground truth annotation of metastases shown via black border. Tiles with feature embeddings larger than 0 are classified as hits and color coded according to their magnitude, with red mapped to strong metastasis detection. **Right:** Detail of metastases at zoom level 2 overlaid with classification output of our proposed approach. Here, the output of all tested tiles are shown, with the feature embeddings mapped to a blue-white-red colormap, with blue and red mapped to strong healthy and metastasis detection, respectively. Tiles without color were not included when randomly selecting tiles for inference.