On the Benefits of Region of Interest Detection for Whole Slide Image Classification

Sena Korkut, Cihan Erkan, and Selim Aksoy

Department of Computer Engineering, Bilkent University, Ankara, 06800, Turkey

ABSTRACT

Whole slide image (WSI) classification methods typically use fixed-size patches that are processed separately and are aggregated for the final slide-level prediction. Image segmentation methods are designed to obtain a delineation of specific tissue types. These two tasks are usually studied independently. The aim of this work is to investigate the effect of region of interest (ROI) detection as a preliminary step for WSI classification. First, we process each WSI by using a pixel-level classifier that provides a binary segmentation mask for potentially important ROIs. We evaluate both single-resolution models that process each magnification independently and multi-resolution models that simultaneously incorporate contextual information and local details. Then, we compare the WSI classification performances of patch-based models when the patches used for both training and testing are extracted from the whole image and when they are sampled from only within the detected ROIs. The experiments using a binary classification setting for breast histopathology slides as benign vs. malignant show that the classifier that uses the patches sampled from the whole image achieves an F1 score of 0.68 whereas the classifiers that use patches sampled from the ROI detection results produced by the single- and multi-resolution models obtain scores between 0.75 and 0.83.

Keywords: Digital pathology, breast histopathology, region of interest detection, whole slide image classification, multi-resolution image analysis

1. INTRODUCTION

Whole slide images (WSIs) that are digitized biopsy slides contain billions of pixels due to their high resolution. This level of detail enables pathologists to observe both the contextual patterns and the individual characteristics of cancer cells. For example, a pathologist may interpret the general pattern of cell groups, and then zoom in to regions of interest (ROIs) to make a more detailed analysis. The combination of these observations allows pathologists to evaluate the existence of cancer and its malignancy level.

The ROIs are defined as regions that are identified to be diagnostically relevant by human experts. For the example case of breast histopathology, the ROIs refer to variations in the tissue content that correspond to proliferative changes in ductal or lobular structures. Localizing and diagnosing these changes necessitate the exploitation of information from different magnifications in the decision process. Furthermore, examination of the viewing behavior of different pathologists shows that they do not necessarily evaluate the whole slide in detail and make their decision by focusing on only some parts of the image.[1]

Advances in deep learning methods on image detection, segmentation, and classification tasks have also found widespread application in computational pathology. However, most studies on histopathological image analysis suffer from the limitations of using WSIs because processing huge images requires high computational power and developing slide-level methods is a complicated task when the diagnostically relevant portions of the image, i.e., the ROIs, occupy only small areas. Hence, most studies focus on manually selected well-defined ROIs compared to the large and complex WSIs. Even though ROI classification can be beneficial for some cases, extending these methods to slide level is not always straightforward.

To process massive WSIs with limited memory resources, slide-level methods usually split an image into fixed-size patches and process each patch separately before aggregating their outputs.[2] While some studies use the sliding window approach to generate the patches, some use random sampling to reduce the computational cost.

Send correspondence to S.A.: E-mail: saksoy@cs.bilkent.edu.tr, Telephone: +90 (312) 2903405

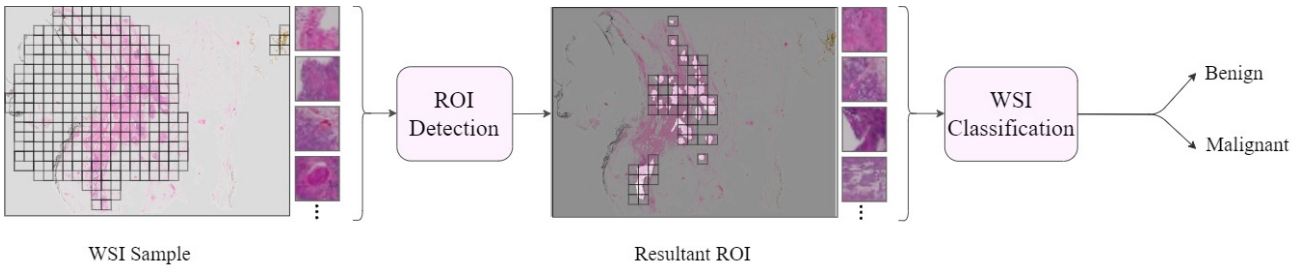

Figure 1. The proposed WSI classification pipeline. Patches sampled from the output of the ROI detection algorithm are input to the WSI classifier that makes the slide-level diagnosis as benign or malignant.

However, not all such patches are equally important for the diagnosis. Furthermore, the contextual information is lost when the patches are treated individually. Some recent work on image semantic segmentation aim to obtain a delineation of specific tissue types by using multi-resolution processing.[3] However, image segmentation and classification tasks are typically studied independently.

In this paper, we study the effect of ROI detection as a preliminary step for WSI classification. Previous work has also shown that fusing ROI detection and patch classification outputs improves the slide-level accuracy.[4] Here, first, we process each WSI by using a pixel-level classifier that provides a binary segmentation for ROIs. We evaluate both single-resolution models that are based on the U-Net[5] architecture and multi-resolution models such as HookNet[3] that simultaneously incorporate contextual information and local details. Then, we compare the WSI classification performance of patch-based models when the patches are extracted from the whole image and when they are sampled from only within the ROIs. In the rest of the paper, we describe the details of the ROI detection methodology and present the experimental results for both ROI detection and WSI classification.

2. METHODOLOGY

The focus of this study is to improve the slide-level classification for breast histopathology images by employing ROI detection as an initial step. This preliminary step aims to imitate the viewing behavior of the pathologist by using multiple magnifications. Given the resultant ROIs identified by the ROI detector, the experiments for WSI classification use convolutional neural network (CNN) models that adopt a patch-based approach by selecting the fixed-size patches from these ROIs. The whole pipeline is shown in Figure 1. The ROI detection step uses binary ROI masks and the WSI classification step uses slide-level diagnostic labels for training.

2.1 ROI detection

The ultimate goal of this step is to produce a slide-wide pixel-level segmentation map for detecting ROIs. Studies centered on cancer detection often adopt an encoder-decoder architecture called U-Net.[5] We also use U-Net as our leading network architecture for both single- and multi-resolution experiments.

Single-resolution architecture. For single-resolution experiments, we use U-net models independently trained by using images at 2.5× and 5× magnifications. The network architecture starts with the encoder path where each block contains two 3×3 convolution operations followed by batch normalization and max pooling layers having a 2×2 down-sampling factor. This process is repeated four times before continuing with the decoder path where multi-channel local information maps acquired from the encoder path are up-sampled by up-convolutional operations. Then, the corresponding feature maps are concatenated with the up-sampled features, followed by two convolutional operations. This process is repeated four times before the last 1×1 convolution, which reduces the channel size to one, producing the final segmentation map.

Multi-resolution architecture. Using U-Net[5] as the leading backbone network, we introduce two different pipelines to process a WSI at multiple magnification levels. The pyramidal pipeline structure considers different resolutions independently and progressively extracts patches from lower resolutions while producing the final segmentation map at the highest resolution. The second structure is based on HookNet,[3] where the higher resolution outputs the final result by obtaining extra information from a larger context in a lower resolution.

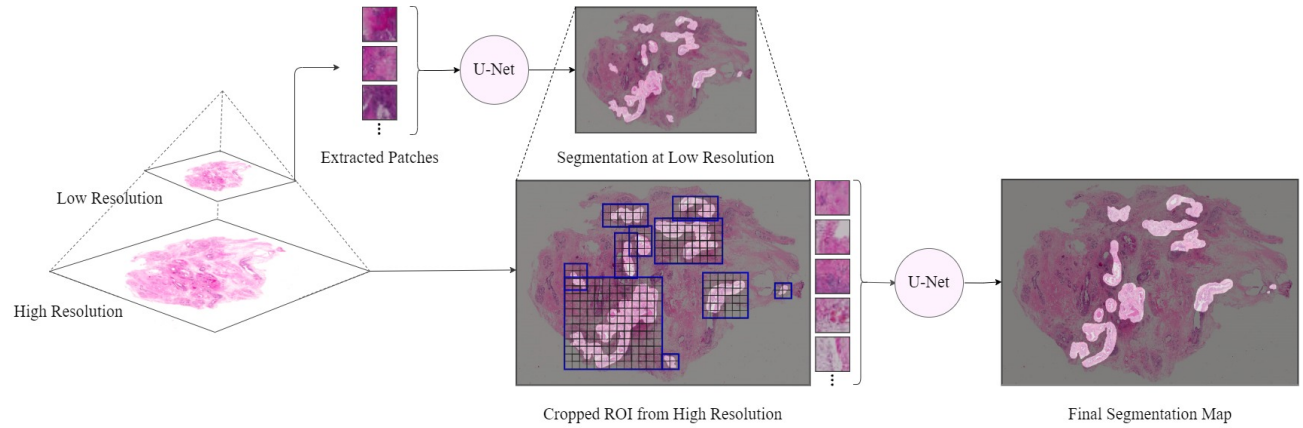

Figure 2. Illustration of the multi-resolution pyramidal pipeline. Patches from low-resolution slide are extracted. The results are cropped from a higher resolution, with a buffer around the bounding box of each potential ROI. The relevant patches are given to another model trained at a higher resolution to produce the final result.

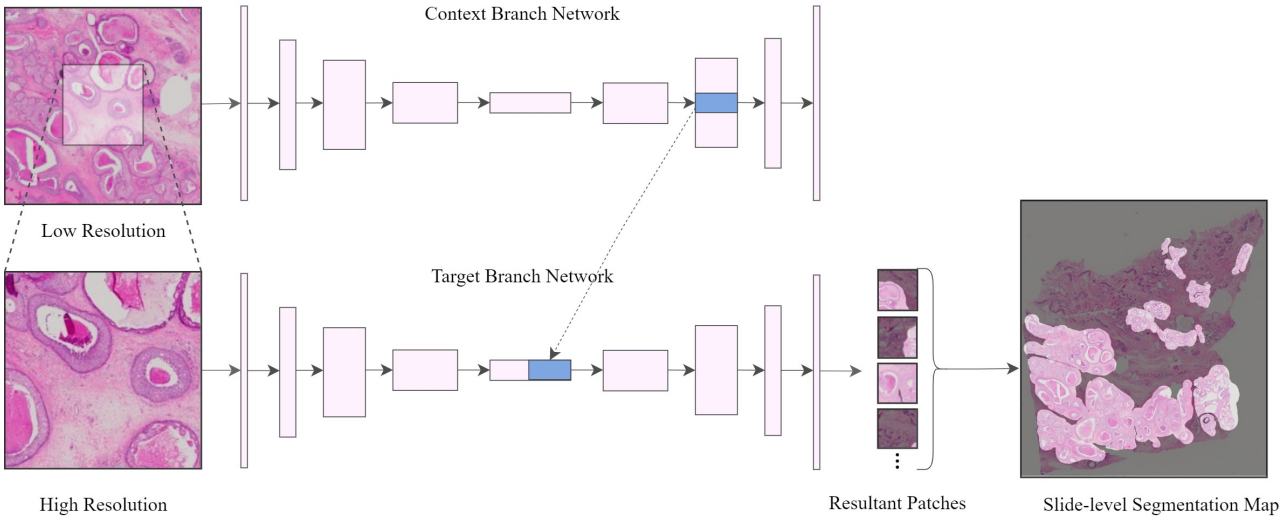

Figure 3. Illustration of the multi-resolution HookNet architecture. Same-sized patches from low and high resolutions are input to two branches resembling U-Net. Considering the resolution, feature vectors from the context branch are "hooked" to the bottleneck layer of the target branch. The final result is generated by processing all patches by the target branch.

- The *pyramidal pipeline* starts by processing the WSI by using a U-Net model that is trained at the lowest resolution. The model outputs a pixel-wise probability map for diagnostically relevant regions. A mask is generated for the slide by thresholding this probability map. Then, connected components of this mask are cropped from a higher resolution image where a buffer is used around the component before cropping. The cropped ROIs are then processed by a different U-Net model that is trained at that resolution. Thresholding of the following output probability map produces a finer ROI mask. This process can continue at higher resolutions but we use only two magnifications, 2.5× and 5×, in the experiments as illustrated in Figure 2.

- *HookNet* is a convolutional model that aims to process patches extracted from different resolutions. It consists of two branches with the same architecture based on the U-net model. The first branch is called the context branch that aims to extract contextual information from low-resolution patches with a large field of view. The second branch is called the target branch that aims to obtain fine-grained details from high-resolution patches with a smaller field of view. Same-sized patches from low and high resolutions are input to HookNet's context and target branches, respectively. Context branch patches have a wider

field of view, and target branch patches represent the center of each patch in a higher resolution. The crucial process of this structure is the part where the feature maps from the context branch are cropped and are concatenated with the feature maps of the target branch, called the "hooking" mechanism. After combining both contextual and fine-grained information into the target branch, the final segmentation map is constructed. The model is illustrated in Figure 3.

2.2 WSI classification

We follow a two-step approach for WSI classification. The first step uses a patch classifier and the second step uses a majority voting-like decision among the patches to obtain the slide-level prediction.[4] The patch classifier is trained on 256×256 pixel patches sampled from the WSIs where the slide-level diagnoses are used as weak labels for the corresponding patches. Patch features used during classification are extracted by using an ImageNet pre-trained ResNet-based feature extractor, and patch classification is done using a multilayer perceptron (MLP) trained using the weak labels.

During inference, the malignancy probability for each patch is calculated using this classifier, and the slide-level diagnosis is obtained as malignant if the average malignancy probability of the patches exceeds a threshold. The slide is classified as benign otherwise. We use this relatively simple approach and the binary diagnosis setting for WSI classification because the dataset used in the experiments does not have sufficient number of slides for training more complex models in a multi-class setting. This process is evaluated when the patches are extracted from the whole image and when they are sampled from only within the detected ROIs during both training and testing.

3. EXPERIMENTS

3.1 Dataset

The dataset was collected from Hacettepe University, Department of Pathology archives. It contains 98 breast WSIs digitized from haematoxylin and eosin-stained specimens of 81 patients. They were scanned by an Olympus slide scanner at $40\times$ magnification and 1376 ROIs were delineated by pathologists in free form. ROI-level annotations are gathered into four diagnostic classes: benign (including samples containing non-proliferative changes, apocrine metaplasia, usual ductal hyperplasia, columnar cell hyperplasia, flat epithelial hyperplasia, and intraductal papilloma without atypia), atypia (including samples containing atypical ductal hyperplasia, atypical lobular hyperplasia, and intraductal papilloma with atypia), in situ carcinoma (including both ductal carcinoma in situ and lobular carcinoma in situ), and invasive carcinoma. Each slide can include multiple labels as ROIs with different classes can reside in the same slide. The final slide-level labels are formed by choosing the most severe diagnostic class from the corresponding multi-label sets. In this work, the ROI detection problem is studied by using binary pixel-level labels as ROI vs. background, and the WSI classification problem is studied by using binary slide-level labels as benign vs. malignant (atypia, in situ carcinoma, invasive).

The specimens in the collection were acquired at various points in time; hence, they were processed by different staining protocols. To reduce the variations among the samples, we perform stain normalization as a preprocessing step. Furthermore, the dataset is split into four folds, by considering each fold's class distribution and ensuring that no two WSIs from the same patient fall into different folds. Table 1 shows the class distribution of the folds.

3.2 Experimental setup

Both the ROI detection and the WSI classification experiments use two folds for training and one fold each for the validation and test sets. This setup is repeated four times so that each slide is used once for testing. We report the average values and confidence intervals for the performance metrics using these cross-validation experiments.

Before training the ROI detection models, all slides are divided into 1024×1024 pixel patches with a 10% overlap. Background patches are eliminated using a threshold on saturation values. In the pyramidal pipeline experiments, the resultant ROIs from the lower magnification are enlarged with a 10% buffer on all sides before cropping the corresponding locations from the higher magnification.

Table 1. Slide-level and ROI-level class distribution of the four folds in the dataset.

		Benign	Atypia	In Situ	Invasive	Total
Slide	Fold 1	9	2	7	7	25
	Fold 2	6	6	4	7	23
	Fold 3	10	0	7	7	24
	Fold 4	7	4	8	7	26
	Total	32	12	26	28	98
ROI	Fold 1	144	35	95	56	330
	Fold 2	147	38	128	59	372
	Fold 3	140	34	102	58	334
	Fold 4	151	39	90	60	340
	Total	582	146	415	233	1376

Table 2. Quantitative results for ROI detection for single- and multi-resolution models.

		Magnification(s)	λ	Precision	Recall	F1
Single-resolution	U-Net	2.5×	-	0.37 ± 0.06	0.62 ± 0.05	0.46 ± 0.03
	U-Net	5×	-	0.34 ± 0.06	0.73 ± 0.05	0.47 ± 0.05
Multi-resolution	Pyramidal	2.5×, 5×	-	0.33 ± 0.05	0.71 ± 0.05	0.45 ± 0.05
	HookNet	2.5×, 5×	1	0.30 ± 0.05	0.75 ± 0.05	0.43 ± 0.05
	HookNet	2.5×, 5×	0.75	0.31 ± 0.03	0.72 ± 0.03	0.43 ± 0.03
	HookNet	2.5×, 5×	0.50	0.24 ± 0.05	0.83 ± 0.04	0.37 ± 0.04

For ROI detection experiments, we use the Adam optimizer with an initial learning rate of 10^{-4} and apply exponential learning rate decay with a value of 0.98. We also use a weight decay of 10^{-6} and a dropout rate of 0.4 after the downsampling layers of U-Net's encoder path. In HookNet experiments, losses of context and target branches are combined by using the function $L = \lambda L_{target} + (1 - \lambda)L_{context}$, where λ represents the relative importance given to the loss of target branch L_{target} and context branch $L_{context}$.[3] We evaluate three different values of λ. Due to class imbalance in the dataset, we employ the focal Tversky loss[6] for all models, using the parameters α, β, γ with the values 0.7, 0.3, and 1.25, respectively. The models are updated with a batch size of 2 patches and are trained for 20 epochs.

For WSI classification experiments, we train the patch classifiers using the cross entropy loss and optimize their parameters using stochastic gradient descent with a learning rate of 0.02 and with a weight decay of 10^{-4}. For the inference phase, we pick malignancy probability thresholds based on their performance on the validation set. For training and inference, we use only the patches classified as ROI by the corresponding ROI detection model. We also conduct experiments without any ROI detection where all foreground patches are used for both training and inference.

3.3 Results

For each combination in the cross-validation experiments, two folds are assigned as training sets, and the other two are assigned as validation and test sets such that all folds are used once for the evaluation of the models. Pixel-level precision and recall are computed for each slide and their mean values together with 95% confidence intervals computed from the folds are obtained for the evaluation of ROI detection. We also compute F1 scores from the resulting mean precision and recall values.

Table 2 summarizes the quantitative performances for the ROI detection step for all single- and multi-resolution experiments. All models perform similarly in terms of F1 scores. The performance for the HookNet architecture decreases with decreasing λ value, indicating that the contribution of the target branch that uses the higher resolution image is important in the final decision. The single-resolution models perform slightly better than the multi-resolution models, suggesting that a more detailed tuning of the parameters of the latter models is needed.

Table 3. Quantitative results for WSI classification when the patches are extracted from the whole slide (Non-ROI) and when the patches are sampled from the ROIs detected by different models.

ROI detection model	Accuracy	Precision	Recall	F1
Non-ROI	0.59 ± 0.14	0.70 ± 0.07	0.67 ± 0.17	0.68 ± 0.12
U-Net (2.5×)	0.65 ± 0.05	0.73 ± 0.09	0.78 ± 0.12	0.75 ± 0.03
U-Net (5×)	0.75 ± 0.01	0.80 ± 0.12	0.87 ± 0.17	0.82 ± 0.04
Pyramidal pipeline	0.76 ± 0.03	0.77 ± 0.04	0.91 ± 0.10	0.83 ± 0.03
HookNet ($\lambda = 1$)	0.73 ± 0.06	0.78 ± 0.11	0.87 ± 0.09	0.81 ± 0.04

The results also show that all ROI detection methods perform poorly in terms of precision, but manage to predict positively labeled pixels relatively well, as indicated in the recall scores. This is due to the difference between the actual content of the slides and the particular masks used for training. Each slide contains several regions that correspond to ductal or lobular structures. However, since the focus of the annotations is to mark the diagnostically significant proliferative changes in these regions, the ROI masks used for training and evaluation do not contain such structures that can be considered as normal. We actually have some additional annotations that include an extra class named *normal* where breast ducts are mainly located. However, these extra markings are not included in the masks used during training and evaluation. Visual inspection of the results show that the ROI detection models tend to attend to nuclei-dense regions and predict the regions considered as normal as well. This is consistent with the high correlation between the regions viewed by the pathologists and epithelium-rich regions in the slides as observed in eye tracking studies.[7] Figure 4 illustrates examples for pathologists' annotations, masks used for training, and output from one of the ROI detection models (HookNet, with $\lambda = 1$). The predictions for normal regions are counted as false positives in the quantitative performance evaluation in Table 2.

A factor that can be considered as a source for decrease in the recall performance is the tendency of the models to leave inner areas of the ductal regions unmarked. Figure 5 shows additional outputs for some ROI detection models along with the masks used in training. The resulting ROIs are generally similar to the reference markings, yet their shapes do not fully match. While the reference masks include more generic and smoother boundaries of ROIs, the predictions are more detailed at pixel-level and are more complex in their shapes. The predictions also often include only the epithelial regions at the outer boundaries of the ducts because the inner details of the ductal regions such as the empty areas, secretion, and necrosis have a high visual similarity to the tissue structures marked as background in the training masks. Moreover, we also observe that the predictions are more accurate for the ROIs that are individually labeled as *benign* or *in situ* as a consequence of their visually consistent structure and frequent representation in the dataset compared to ROIs labeled as *atypia* that occupy relatively smaller areas and rarely occur in the dataset.

Table 3 presents the quantitative performances for WSI classification when the patches used for both training and testing are extracted from the whole slide and when they are sampled from only within the ROIs detected by using different models. The results show that slide-level classification performance can be significantly improved by removing the uninformative patches with the support of the ROI detection step. Both the classification accuracy and the F1 score are improved with all ROI detection models compared to the baseline performance of using no ROI information. Among the ROI detection approaches, the multi-resolution pyramidal pipeline is the best performer in terms of both accuracy and F1 score.

4. CONCLUSION

We studied the effect of ROI detection as a preliminary step for WSI classification. First, we evaluated both single-resolution and multi-resolution models that were trained on pixel-level binary masks for ROI vs. background classification. Then, we compared the performance for patch-based classification of breast histopathology slides as benign vs. malignant when the patches used for both training and testing were extracted from the whole image and when they were sampled from only within the detected ROIs. The results showed that using the predicted ROIs for sampling the patches produced significantly better results in terms of all performance metrics. We

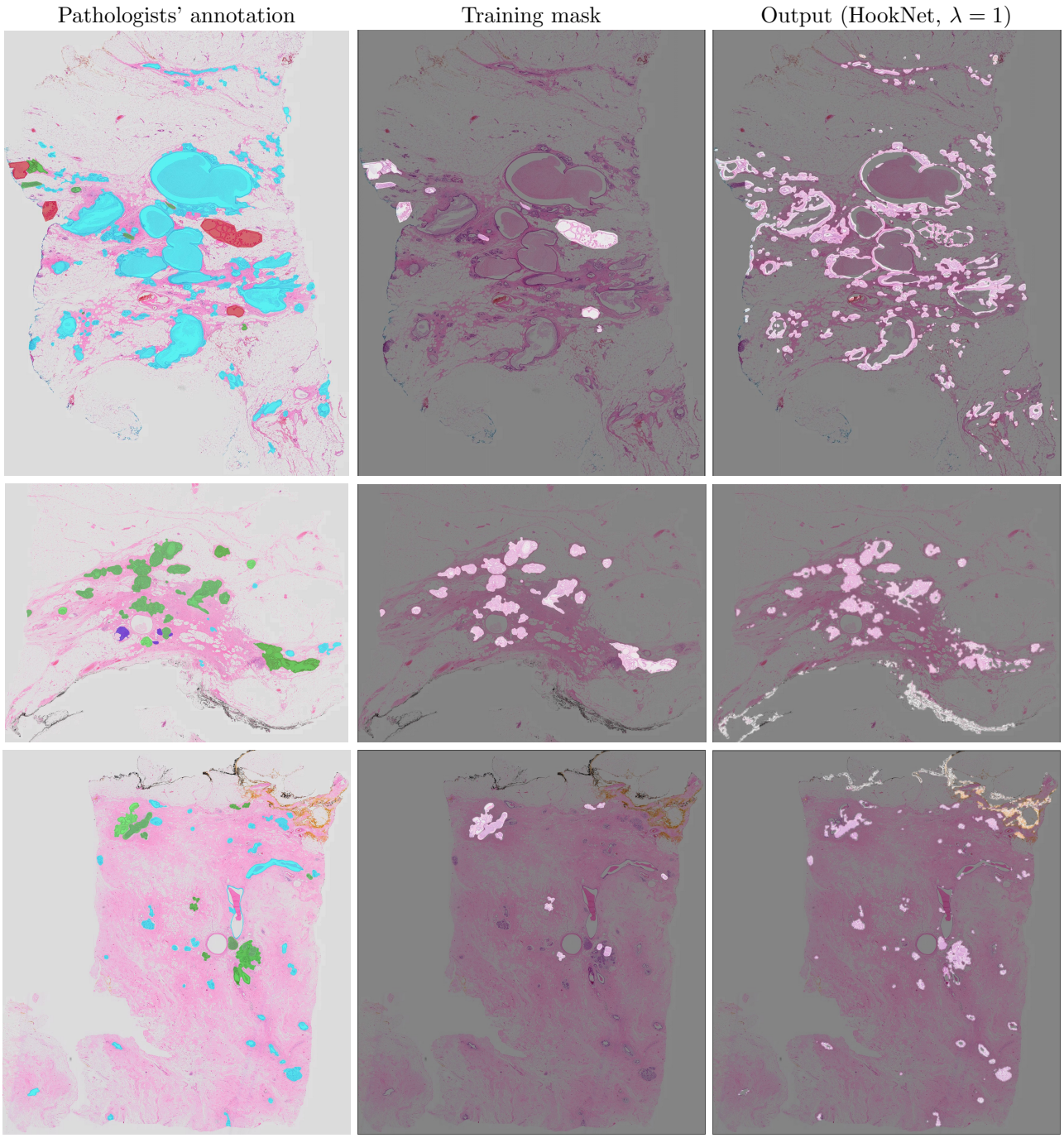

Figure 4. Example slides with an overlay of pathologists' ROI annotations (left), binary reference masks used for training and evaluating the ROI detection models (middle), and predictions by the HookNet model (right). Light blue ROIs in the annotations represent the *normal* category, red ROIs represent *apocrine metaplasia* (*benign*), green ROIs represent *usual ductal hyperplasia* and *columnar cell change hyperplasia* (also *benign*), and dark blue ROIs represent *lobular carcinoma in situ*. The ROI detection model successfully predicts most of the normal regions in the pathologists' extra annotations but these regions are not considered as positive during training or evaluation as they are not included in the masks.

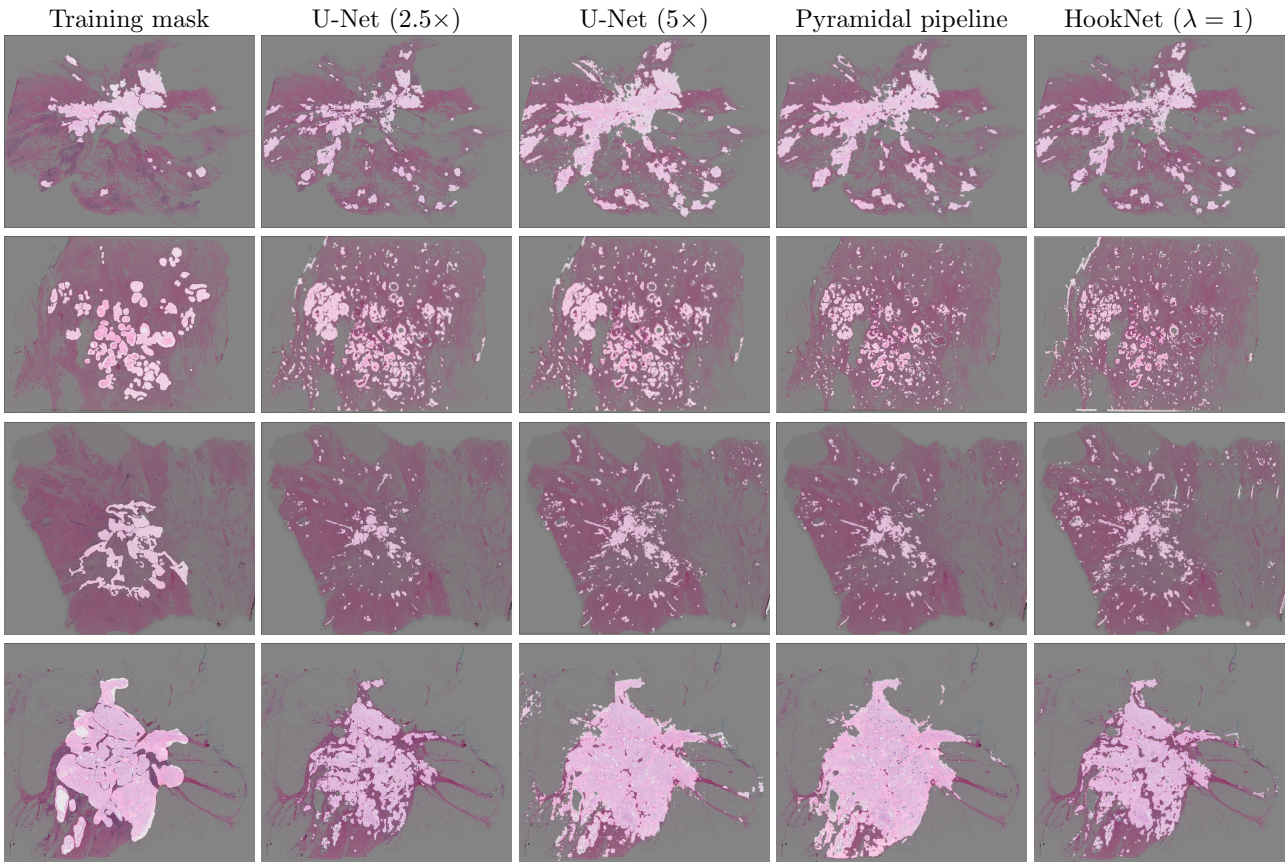

Figure 5. Example outputs from different ROI detection models. The binary masks for ROI detection do not include the regions annotated as *normal*. The first row presents a slide that includes ROIs of *benign*, *atypia*, and *in situ* classes, as well as many *normal* ROIs. The slide in the second row is dominated by *in situ* ROIs, but also contains some ROIs of type *benign*, *invasive*, and *normal*. The third example contains large areas of *in situ* and *invasive* ROIs, but also includes *benign* and *normal* regions. The fourth row includes a slide that mainly consists of *benign* and *atypia* regions.

believe that further enhancements in ROI detection using additional refinement of ROI masks, better parameter tuning, and longer training epochs can lead to even further improvements in WSI classification.

Acknowledgment

This work was supported in part by the GEBIP Award from the Turkish Academy of Sciences.

REFERENCES

[1] Mercan, C., Aksoy, S., Mercan, E., Shapiro, L. G., Weaver, D. L., and Elmore, J. G., "Multi-instance multi-label learning for multi-class classification of whole slide breast histopathology images," *IEEE Transactions on Medical Imaging* **37**, 316–325 (January 2018).

[2] Lu, M. Y., Williamson, D. F., Chen, T. Y., Chen, R. J., Barbieri, M., and Mahmood, F., "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomedical Engineering* **5**(6), 555–570 (2021).

[3] van Rijthoven, M., Balkenhol, M., Silina, K., van der Laak, J., and Ciompi, F., "HookNet: Multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images," *Medical Image Analysis* **68**, 101890 (2021).

[4] Gecer, B., Aksoy, S., Mercan, E., Shapiro, L. G., Weaver, D. L., and Elmore, J. G., "Detection and classification of cancer in whole slide breast histopathology images using deep convolutional networks," *Pattern Recognition* **84**, 345–356 (December 2018).

[5] Ronneberger, O., Fischer, P., and Brox, T., "U-Net: Convolutional networks for biomedical image segmentation," (2015). arXiv:1505.04597.

[6] Abraham, N. and Khan, N. M., "A novel focal Tversky loss function with improved attention U-Net for lesion segmentation," (2018). arXiv:1810.07842.

[7] Brunye, T. T., Carney, P. A., Allison, K. H., Shapiro, L. G., Weaver, D. L., and Elmore, J. G., "Eye movements as an index of pathologist visual expertise: A pilot study," *PLoS ONE* **9**(8) (2014).