Y-Net: Joint Segmentation and Classification for Diagnosis of Breast Biopsy Images

Sachin Mehta[1], Ezgi Mercan[1], Jamen Bartlett[2], Donald Weaver[2], Joann G. Elmore[1], and Linda Shapiro[1]

[1] University of Washington, Seattle WA 98195, USA
[2] University of California, Los Angeles CA 90095, USA
[3] University of Vermont, Burlington VT 05405, USA
{sacmehta, ezgi, shapiro}@cs.washington.edu {jamen.bartlett, donald.weaver}@uvmhealth.org jelmore@mednet.ucla.edu

Abstract. In this paper, we introduce a conceptually simple network for generating discriminative tissue-level segmentation masks for the purpose of breast cancer diagnosis. Our method efficiently segments different types of tissues in breast biopsy images while simultaneously predicting a discriminative map for identifying important areas in an image. Our network, Y-Net, extends and generalizes U-Net by adding a parallel branch for discriminative map generation and by supporting convolutional block modularity, which allows the user to adjust network efficiency without altering the network topology. Y-Net delivers state-of-the-art segmentation accuracy while learning 6.6× fewer parameters than its closest competitors. The addition of descriptive power from Y-Net's discriminative segmentation masks improve diagnostic classification accuracy by 7% over state-of-the-art methods for diagnostic classification.

1 Introduction

Annually, millions of women depend on pathologists' interpretive accuracy to determine whether their breast biopsies are benign or malignant [4]. Diagnostic errors are alarmingly frequent, lead to incorrect treatment recommendations, and can cause significant patient harm [2]. Pathology as a field has been slow to move into the digital age, but in April 2017, the FDA authorized the marketing of the Philips IntelliSite Pathology Solution (PIPS), the first whole slide imaging system for interpreting digital surgical pathology slides on the basis of biopsy tissue samples, thus changing the landscape[4].

Convolutional neural networks (CNNs) produce state-of-the-art results in natural [12, 6] and biomedical classification and segmentation [8, 11] tasks. Training CNNs directly on whole slide images (WSIs) is difficult due to their massive size. Sliding-window-based approaches for classifying [8, 5] and segmenting [11, 10] medical images have shown promising results. Segmentation and classification are usually separate steps in automated diagnosis systems.

[4] https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm552742.htm

Segmentation-based methods consider tissue structure, such as size and distribution, to help inform class boundary decisions. However, these segmentation methods suffer from two major drawbacks. First, labeled data is scarce because the labeling of biopsy images is time-consuming and must be done by domain experts. Second, segmentation-based approaches are not able to weigh the importance of different tissue types. The latter limitation is particularly concerning in biopsy images, because not every tissue type in biopsy images is relevant for cancer detection. On the other hand, though classification-based methods fail to provide structure- and tissue-level information, they can identify regions of interest inside the images that should be used for further analysis.

In this paper, we combine the two different methods, segmentation and classification, and introduce a new network called Y-Net that simultaneously generates a tissue-level segmentation mask and a discriminative (or saliency) map. Y-Net generalizes the U-Net network [11], a well-known segmentation network for biomedical images. Y-net includes a *plug-and-play* functionality that enables the use of different types of convolutional blocks without changing the network topology, allowing users to more easily explore the space of networks and choose more efficient networks. For example, Y-Net delivers the same segmentation performance as that of [10] while learning 6.6× fewer parameters. Furthermore, the discriminative tissue-level segmentation masks produced using Y-Net provide powerful features for diagnosis. Our results suggest that Y-Net is 7% more accurate than state-of-the-art segmentation and saliency-based methods [10, 5].

Statement of problem: The problem we wish to solve is the simultaneous segmentation and diagnosis of whole slide breast cancer biopsy images. For this task, we used the breast biopsy dataset in [2, 10] that consists of 240 whole slide breast biopsy images with heamatoxylin and eosin (H&E) staining. A total of 87 pathologists diagnosed a randomly assigned subset of 60 slides into four diagnostic categories (benign, atypia, ductal carcinoma *in situ*, and invasive cancer), producing an average of 22 diagnostic labels per case. Then, each slide was carefully interpreted by a panel of three expert pathologists to assign a consensus diagnosis for each slide that we take to be the gold standard ground truth. Furthermore, the pathologists have marked 428 regions of interest (ROIs) on these slides that helped with diagnosis and a subset of 58 of these ROIs have been hand segmented by a pathology fellow into eight different tissue classifications: *background, benign epithelium, malignant epithelium, normal stroma, desmoplastic stroma, secretion, blood,* and *necrosis*. The average size of these ROIs is $10,000 \times 12,000$. We use these 428 ROIs for our data set.

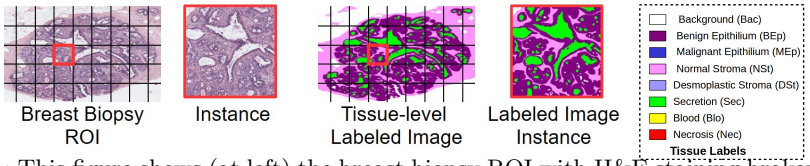


Fig. 1: This figure shows (at left) the breast biopsy ROI with H&E staining broken into multiple instances with one instance enlarged to show more detail. On the right are the pixel-wise tissue-level labelings of the ROI and the instance.

In this work, we break each ROI into a set (or bag) of equal size patches that we will call *instances*, as shown in Figure 1. Each ROI X has a known groundtruth diagnostic label Z. There are no separate diagnostic labels for the instances; they all have the same groundtruth label Z, but some of them contribute to the diagnosis of Z and others do not. Our system, therefore, will learn the *discriminativeness* of each instance during its analysis. Furthermore, each pixel of each of these instances has a known tissue classification into one of the eight categories; tissue classification must be learned from the groundtruth ROIs. Using the groundtruth diagnostic labels Z of the ROIs and the groundtruth tissue labels Y from the 58 labeled ROIs, our goal is to build a classification system that can input a ROI, perform simultaneous segmentation and classification, and output a diagnosis. Our system, once trained, can be easily applied to WSIs.

Related work: Biomedical images are difficult to classify and segment, because their anatomical structures vary in shape and size. CNNs, by virtue of their representational power and capacity for capturing structural information, have made such classification and segmentation tasks possible [11, 8]. The segmentation-based method in [10] and saliency map-based method in [5] are most similar to our work. Mehta *et al.* [10] developed a CNN-based method for segmenting breast biopsy images that produces a tissue-level segmentation mask for each WSI. The histogram features they extracted from the segmentation masks were used for diagnostic classification. Geçer [5] proposed a saliency-based method for diagnosing cancer in breast biopsy images that identified relevant regions in breast biopsy WSIs to be used for diagnostic classification. Our main contribution in this paper is a method for *joint learning* of both segmentation and classification. Our experiments show that joint learning improves diagnostic accuracy.

2 A System for Joint Segmentation and Classification

Our system (Figure 2) is given an ROI from a breast biopsy WSI and breaks it into instances that are fed into Y-Net. Y-Net produces two different outputs: an instance-level segmentation mask and an instance-level probability map. The instance-level segmentation masks have, for each instance, the predicted labels of the eight different tissue types. These are combined to produce a segmentation mask for the whole ROI. The instance-level probability map contains (for every pixel) the maximum value of the probability of that instance being in one of the four diagnostic categories. This map is thresholded to binary and combined with the segmentation mask to produce the discriminative segmentation mask.

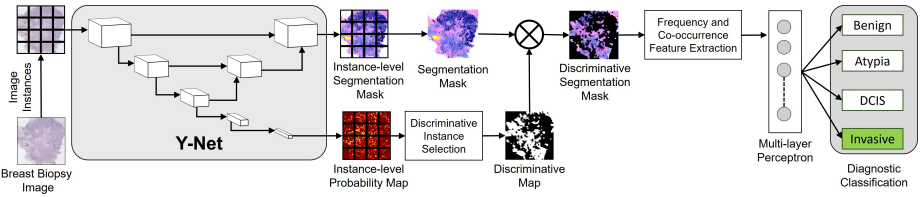


Fig. 2: Overview of our method for detecting breast cancer.

.

A multi-layer perceptron then uses the frequency and co-occurrence features extracted from the final mask to predict the cancer diagnosis.

2.1 Y-Net Architecture

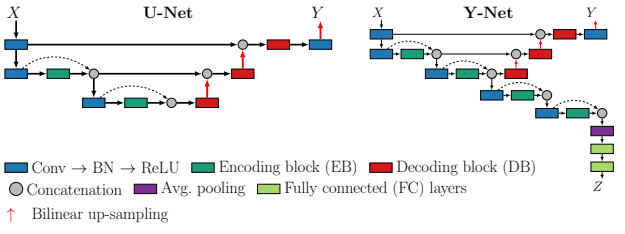
Y-Net is conceptually simple and generalizes U-Net [11] to joint segmentation and classification tasks. U-Net outputs a single segmentation mask. Y-Net adds a second branch that outputs the classification label. The classification output is distinct from the segmentation output and requires feature extraction at low spatial resolutions. We first briefly review the U-Net architecture and then introduce the key elements in the Y-Net architecture.

U-Net: U-Net is composed of two networks: (1) encoding network and (2) decoding network. The encoding network can be viewed as a stack of encoding and down-sampling blocks. The encoding blocks learn input representations; down-sampling helps the network learn scale invariance. Spatial information is lost in both convolutional and down-sampling operations. The decoder can be viewed as a stack of up-sampling and decoding blocks. The up-sampling blocks help in inverting the loss of spatial resolution, while the decoding blocks help the network to compensate for the loss of spatial information in the encoder. U-Net introduces skip-connections between the encoder and the decoder, which enables the encoder and the decoder to share information.

Y-Net: Y-Net (Figure 3a) adopts a two-stage procedure. The first stage outputs the instance-level segmentation mask, as U-Net does, while the second stage adds a parallel branch that outputs the instance-level classification label. In spirit, our approach follows Mask-RCNN [7] which jointly learns the segmentation and classification of natural images. Unlike Mask-RCNN, Y-Net is fully convolutional; that is, Y-Net does not have any region proposal network. Furthermore, training Y-Net is different from training Mask-RCNN, because Mask-RCNN is trained with object-level segmentations and classification labels. Our system has diagnostic labels for entire ROIs, but not for the instance-level.

Y-Net differs from the U-Net in the following aspects:

Abstract representation of encoding and decoding blocks: At each spatial level, U-Net uses the same convolutional block (a stack of convolutional layers) in both the encoder and the decoder. Instead, Y-Net abstracts this representation and represents convolutional blocks as general encoding and decoding blocks that



(a) U-Net vs. Y-Net

Spatial Level	Block Name	# Blocks (depth)	# Channels (width)
1	Conv	1	16
2	Conv	1	w
	EB	2	$2w$
3	Conv	1	$2w$
	EB	d	$4w$
4	Conv	1	w
	EB	2	$2w$
5	Conv	1	$w/2$
	EB	2	w
6	Avg. Pool	1	w
	FC	1	64
	FC	1	Classes

(b) Encoding Network

Fig. 3: (a) Comparison between U-Net and Y-Net architectures. (b) The encoding network architecture used in (a). U-Net in (a) is a generalized version of U-Net [11].

can be used anywhere, and are thus not forced to be the same at each spatial level. Representing Y-Net in such a modular form provides it a *plug-and-play* functionality and therefore enables the user to try different convolutional block designs without changing the network topology.

Width and depth multipliers: Larger CNN architectures tend to perform better than smaller architectures. We introduce two hyper-parameters, a width multiplier w and a depth-multiplier d, that allow us to vary the size of the network. These parameters allow Y-Net to span the network space from smaller to larger networks, allowing identification of better network structures.

Sharing features: While U-Net has skip-connections between the encoding and decoding stages, Y-net adds a skip-connection between the first and last encoding block at the same spatial resolution in the encoder, as shown in Figure 3 with a dashed arrow, to help improve segmentation.

Implementation details: The encoding network in Y-Net (Figure 3a) consists of the repeated application of the encoding blocks and 3×3 convolutional layers with a stride of 2 for down-sampling operations, except for the first layer which is a 7×7 standard convolution layer with a stride of 2. Similarly, the decoding network in Y-Net consists of the repeated application of the decoding blocks and bilinear up-sampling for up-sampling operations. We first train Y-Net for segmentation and then attach the remaining encoding network (spatial levels 4, 5 and 6 in Figure 3b) to jointly train for segmentation and classification. We define a multi-task loss on each instance as $\mathcal{L} = \mathcal{L}_{seg} + \mathcal{L}_{cls}$, where $\mathcal{L}_{seg}$ and $\mathcal{L}_{cls}$ are the multi-nominal cross-entropy loss functions for the segmentation and classification tasks, respectively. All layers and blocks, except the classification and fully connected (FC) layers, are followed by a batch normalization and ReLU non-linearity. An average pooling layer with adaptive kernel size enables Y-Net to deal with arbitrary image sizes.

2.2 Discriminative Instance Selection

The encoding network in Y-Net generates a C-dimensional output vector $\mathbf{z}$ of real values; C represents the number of diagnostic classes. The real-values in $\mathbf{z}$ are normalized using a softmax function σ to generate another C-dimensional vector $\bar{\mathbf{z}} = \sigma(\mathbf{z})$. It is reasonable to assume that instances with low probability will have low discriminativeness. If $\max(\bar{\mathbf{z}}) > \tau$, then the instance is considered *discriminative*, where τ is the threshold selected using the method in [8].

2.3 Diagnostic Classification

Segmentation masks provide tissue-level information. Since training data with tissue-level information is limited and not all tissue types contribute equally to diagnostic decisions, our system combines the segmentation mask with the discriminative map to obtain a *tissue-level discriminative segmentation mask*. Frequency and co-occurrence histograms are extracted from the discriminative segmentation mask and used to train a multi-layer perceptron (MLP) with 256, 128, 64, and 32 hidden nodes to predict the diagnostic class.

3 Experiments

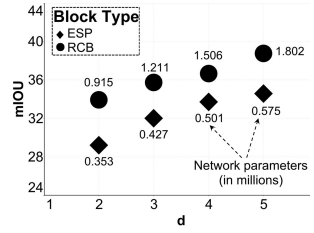
In this section, we first study the effect of the modular design in Y-Net. We then compare the performance of Y-Net with state-of-the-art methods on tissue-level segmentation as well as on diagnostic classification tasks. For evaluation, we used the breast biopsy dataset [2, 10] that consists of 428 ROIs with classification labels and 58 ROIs with tissue-level labels.

3.1 Segmentation Results

We used residual convolutional blocks (RCB) [6] and efficient spatial pyramid blocks (ESP) [9] for encoding and decoding. Based on the success of PSPNet for segmentation [12], we added pyramid spatial pooling (PSP) blocks for decoding.
Training details: We split the 58 ROIs into nearly equal training (# 30) and test (# 28) sets. For training, we extracted 384×384 instances with an overlap of 56 pixels at different image resolutions. We used standard augmentation strategies, such as random flipping, cropping, and resizing, during training. We used a 90:10 ratio for splitting training data into training and validation sets. We trained the network for 100 epochs using SGD with an initial learning rate of 0.0001, decaying the rate by a factor of 2 after 30 epochs. We measured the accuracy of each model using mean Region Intersection over Union (mIOU).
Segmentation studies: Segmentation results are given in Table 1 and Table 2a. We make the following observations:
Feature sharing: (Table 1a) When features were shared between the encoding blocks at the same spatial level, the accuracy of the network improved by about 2% (R2 and R5) with element-wise addition operations and about 4% (R3 and R6) with concatenation operations for both ESP and RCB blocks. The increase in number of parameters due to concatenation operations was not significant.
Network depth: (Table 1b) The value of d was varied from 2 to 5 in Y-Net for both ESP and RCB types of convolutional blocks. The accuracy of the network increased with the depth of the network. When we increased d from 2 to 5, the accuracy improved by about 4% while the network parameters were increased by about 1.6× and 1.9× for Y-Net with ESP and RCB respectively. In the following experiments, we used $d = 5$.

Row	Encoding Block		Decoding Block		Feature Sharing		# Params (in million)	mIOU	# Params (in million)	mIOU
#	ESP	RCB	ESP	RCB PSP	Add	Concat	$w = 64$		$w = 128$	
R1	✓		✓				**0.49**	30.39	**1.95**	35.23
R2	✓		✓		✓		**0.49**	32.12	**1.95**	36.19
R3	✓		✓			✓	0.57	34.58	2.25	38.03
R4		✓		✓			1.72	33.05	6.84	37.93
R5		✓		✓	✓		1.72	36.34	6.84	39.21
R6		✓		✓		✓	1.81	38.75	7.16	40.23
R7	✓			✓		✓	0.69	37.12	2.75	44.03
R8		✓		✓		✓	1.91	**41.96**	7.62	**44.19**

(a) Network width vs. accuracy



(b) Network depth vs. accuracy

Table 1: Ablation studies on Y-Net. In (a), we used $d = 5$. In (b), we used $w = 64$. The experimental settings in (b) were the same as for R3 and R6 in (a).

Network	mIOU	# Params (in million)
Superpixel + SVM [10]	25.8	NA
Badrinarayanan *et al.* [1]	37.6	12.8
Fakhry *et al.* [3]	38.1	12.8
Mehta *et al.* [10]	**44.20**	26.03
YNet (ESP-PSP) - seg	44.03	**2.75**
YNet (RCB-PSP) - seg	44.19	7.62
YNet (ESP-PSP) - joint	43.24	3.91
YNet (RCB-PSP) - joint	43.11	9.11

(a) Segmentation results

Feature Type	Accuracy (in %)
Pathologists (# 44)	70.0
LAB + LBP features [5]	45.0
Segmentation mask [10]	54.5
Saliency map [5]	55.0
Y-Net with different choices	
Segmentation mask	53.25
-background	52.22
-stroma	48.06
Discriminative mask	**62.50**

(b) Diagnostic classification results

Table 2: Comparison with state-of-the-art methods. seg: training Y-Net only for the segmentation task; joint: joint learning for segmentation and classification tasks.

Network width: (Table 1a) When the value of w changed from 64 to 128, the accuracy of Y-Net with ESP (R1-R3) and RCB (R4-R6) increased by about 4%. However, the number of network parameters increased drastically.

PSP as decoding block: (Table 1a) Changing the decoding block from ESP and RCB to PSP helped improve the accuracy by about 3%. This is because the pooling operations in PSP modules helped the network learn better global contextual information. Surprisingly, when the value of w increased from 64 to 128, Y-Net with ESP and PSP delivered accuracies similar to PCB and PSP. This is likely due to the increased number of kernels per branch in the ESP blocks, which helps to learn better representations. Y-Net with ESP blocks learns about 3× fewer parameters and is therefore more efficient.

Joint Training: (Table 2a) Training Y-Net jointly for both classification and segmentation tasks dropped the segmentation accuracy by about 1%. This is likely because we trained the network using an instance-based approach and we did not have classification labels at instance-level.

Comparison with state-of-the-art: (Table 2a) Y-Net outperformed the plain [1] and residual [3] encoder-decoder networks by 7% and 6% respectively. With the same encoding block (RCB) as in [10], Y-Net delivered a similar accuracy while learning 2.85× fewer parameters. We note that Y-Net with ESP and PSP blocks also delivered a similar performance while learning 6.6× fewer parameters than [10] and 2.77× fewer parameters than Y-Net with RCB and PSP blocks. Therefore, the modular architecture of Y-Net allowed us to explore different convolutional blocks with a minimal change in the network topology to find a preferred network design.

3.2 Diagnostic Classification Results

For classification experiments, we split the 428 ROIs in the dataset into almost equal training (# 209) and test (# 219) sets while maintaining the same class distribution across both the sets. We note that the 30 ROIs used for training the segmentation part were part of the training subset during the classification task. The tissue-level segmentation mask and discriminative map were first generated using Y-Net with ESP as encoding blocks and PSP as decoding blocks, which were then used to generate the discriminative segmentation mask. A 44-dimensional feature vector (frequency and co-occurrence histogram) was then

extracted from the discriminative mask. These features were used to train a MLP that classifies the ROI into four diagnoses (benign, atypia, DCIS, and invasive cancer).

A summary of results is given in Table 2b. The classification accuracy improved by about 9% when we used discriminative masks instead of segmentation masks. Our method outperformed state-of-the-art methods that use either the segmentation features [10] or the saliency map [5] by a large margin. Our method's 62.5% accuracy is getting closer to the 70% accuracy of trained pathologists in a study [2]. This suggests that the discriminative segmentation masks generated using Y-Net are powerful.

4 Conclusion

The Y-Net architecture achieved good segmentation and diagnostic classification accuracy on a breast biopsy dataset. Y-Net was able to achieve the same segmentation accuracy as state-of-the-art methods while learning fewer parameters. The features generated using discriminative segmentation masks were shown to be powerful and our method was able to attain higher accuracy than state-of-the-art methods. Though we studied breast biopsy images in this paper, we believe that Y-Net can be extended to other medical imaging tasks.

Acknowledgements: Research reported in this publication was supported by the National Cancer Institute awards R01 CA172343, R01 CA140560, and RO1 CA200690. We would also like to thank NVIDIA Corporation for donating the Titan X Pascal GPU used for this research.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. TPAMI (2017)
2. Elmore et al.: Diagnostic concordance among pathologists interpreting breast biopsy specimens. JAMA (2015)
3. Fakhry, A., Zeng, T., Ji, S.: Residual deconvolutional networks for brain electron microscopy image segmentation. IEEE transactions on medical imaging (2017)
4. Fine, R.E.: Diagnostic Techniques. B. C. Decker, Inc., Ontario, 2nd edn. (2006)
5. Geçer, B.: Detection and classification of breast cancer in whole slide histopathology images using deep convolutional networks. Ph.D. thesis, Bilkent Univ. (2016)
6. He et al.: Deep residual learning for image recognition. In: CVPR (2016)
7. He et al.: Mask r-cnn. In: ICCV (2017)
8. Hou et al.: Patch-based convolutional neural network for whole slide tissue image classification. In: CVPR (2016)
9. Mehta et al.: ESPNet: Efficient Spatial Pyramid of Dilated Convolutions for Semantic Segmentation. arXiv preprint arXiv:1803.06815 (2018)
10. Mehta et al.: Learning to segment breast biopsy whole slide images. WACV (2018)
11. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
12. Zhao et al.: Pyramid scene parsing network. In: CVPR (2017)