Graph Convolutional Networks for Region of Interest Classification in Breast Histopathology

Bulut Aygüneş[a], Selim Aksoy[a], Ramazan Gökberk Cinbiş[b], Kemal Kösemehmetoğlu[c], Sevgen Önder[c], and Ayşegül Üner[c]

[a]Department of Computer Engineering, Bilkent University, Ankara, 06800, Turkey
[b]Department of Computer Engineering, METU, Ankara, 06800, Turkey
[c]Department of Pathology, Hacettepe University, Ankara, 06100, Turkey

ABSTRACT

Deep learning-based approaches have shown highly successful performance in the categorization of digitized biopsy samples. The commonly used setting in these approaches is to employ convolutional neural networks for classification of data sets consisting of images all having the same size. However, the clinical practice in breast histopathology necessitates multi-class categorization of regions of interest (ROI) in biopsy samples where these regions can have arbitrary shapes and sizes. The typical solution to this problem is to aggregate the classification results of fixed-sized patches cropped from these images to obtain image-level classification scores. Another limitation of these approaches is the independent processing of individual patches where the rich contextual information in the complex tissue structures has not yet been sufficiently exploited. We propose a generic methodology to incorporate local inter-patch context through a graph convolution network (GCN) that admits a graph-based ROI representation. The proposed GCN model aims to propagate information over neighboring patches in a progressive manner towards classifying the whole ROI into a diagnostic class. The experiments using a challenging data set for a 4-class ROI-level classification task and comparisons with several baseline approaches show that the proposed model that incorporates the spatial context by using graph convolutional layers performs better than commonly used fusion rules.

Keywords: Digital pathology, breast histopathology, region of interest classification, weakly supervised learning

1. INTRODUCTION

Breast cancer patients can face a variety of clinical actions such as surgery, radiation, or hormonal therapy depending on the diagnosis made by the pathologists for the biopsy samples. Different types of proliferations in the tissue structures carry different risks of progressing into malignancy; thus, the accuracy of the diagnosis in a fine-grained multi-class setting becomes critical.

Histopathological image analysis aims to serve as an important tool for helping pathologists with the diagnostic process. Deep learning-based approaches, in particular convolutional neural networks (CNN), have been shown to be successful in image classification tasks from various domains including digital pathology.[1] As the mainstream CNN architectures for image classification typically require fixed-sized inputs, their common use in the digital pathology domain has also been in the classification of fixed-sized biopsy image patches. For example, commonly used data sets[2,3] include fixed-sized images that are manually selected from biopsy samples with the goal of preparing benchmarks in image classification competitions. The generally studied setting has been to aggregate the classification results of fixed-sized patches cropped from these images to obtain an image-level classification score. Aggregation methods typically include fixed fusion rules such as averaging class probabilities or majority voting.[4–7]

However, whole slide images (WSI) that are obtained by digitizing biopsy slides at high magnification often contain many regions of interest (ROI) that can belong to different diagnostic categories and can carry different levels of relevance for the slide-level diagnosis. Furthermore, the pathologists do not have any restrictions on the

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

Medical Imaging 2020: Digital Pathology, edited by John E. Tomaszewski, Aaron D. Ward, Proc. of SPIE Vol. 11320, 113200K · © 2020 SPIE
CCC code: 1605-7422/20/$21 · doi: 10.1117/12.2550636

ROI size when they evaluate the slides, and can select and study the regions at any size and magnification deemed suitable. Therefore, multi-class classification of arbitrarily sized ROIs appears to be an important problem that serves as a necessary step in the diagnostic process of breast cancer.

We have been studying this problem in a weakly supervised learning perspective where the contributions of the individual patches to the ROI-level diagnosis are not known during training. We proposed a generic feature representation for arbitrarily sized ROIs by using weighted aggregation of the feature representations of fixed-sized patches sampled from these ROIs.[8] Both the patch-level feature representations and the weights were obtained from a convolutional network trained on patches sampled from ROIs in the training data. Similarly, representations like bag-of-words or Fisher encodings are also suitable methods to obtain feature representations for arbitrarily sized ROIs. However, all aforementioned ROI feature representations and aggregation methods based on patch classification results move from the patch level to the ROI level without exploiting the spatial information within the neighborhoods of the patches.

The spatial context formed by individual patches towards their collective contribution to the ROI-level diagnosis remains to be an important detail that has not been studied in earlier work in this domain. In this paper, we propose to incorporate local context through a graph-based ROI representation over a variable number of patches and their spatial proximity relations. More specifically, we formulate the ROI classification problem as a graph classification problem where vertices denote the patches sampled from a given ROI and edges represent the spatial proximity of those patches. The graph structure, therefore, implicitly encodes the spatial relationships across the patches, which can be used to tackle fine-grained ROI classification in a much more holistic manner compared to mainstream patch-classification based approaches.

To realize a classification model for the proposed ROI graph representation, we propose a graph convolutional network (GCN) based classification model. GCN models have been previously shown to be effective in the utilization of the spatial context in visual inputs.[9,10] However, the application of GCNs in the digital pathology domain is limited.[11] Our proposed GCN architecture extracts per-patch representations, propagates information over the neighboring patches in a progressive manner to incorporate the spatial context, and finally, aggregates the resulting patch representations to classify the whole ROI into a diagnostic class. Our experimental results demonstrate the power of the proposed GCN architecture over a number of strong baselines.

In the following, we first introduce the breast pathology data set used in the paper, then, describe the adapted graph convolutional network model, and finally, present the experimental results.

2. DATA SET

We constructed a new data set that currently contains 1,030 ROIs annotated within 78 WSIs that were digitized from haematoxylin and eosin stained specimens belonging to 63 different patients. The specimens were selected from the archives of the Department of Pathology at Hacettepe University based on their slide-level diagnoses. The WSIs were acquired at 40× magnification by using an Olympus slide scanner, resulting in an average image size of 170,000 × 132,000 pixels. The ROIs were annotated by experienced pathologists in free form with no restriction in the sizes and shapes of the image masks. The resulting annotations were collected into 4 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. The class-specific ROI size statistics in Table 1 show a high variation for the samples in the data set.

Table 1. ROI size statistics per diagnostic class in number of pixels at 10× magnification. Rows show the average ROI size, the standard deviation of ROI sizes, and the ratio of the largest ROI size to the smallest one, respectively.

	Benign	Atypia	In Situ	Invasive
Average	$1308K$	$473K$	$2815K$	$12568K$
Standard deviation	$2510K$	$711K$	$4948K$	$17822K$
Max-min ratio	977.2	210.8	941.1	762.5

Table 2. Class distribution of slides and ROIs in training, validation, and test sets. Note that a slide can contain multiple ROIs corresponding to different diagnostic labels, resulting in a multi-label setting for each slide. Thus, the numbers of slides for each diagnostic class in the table do not sum up to the total number of slides for a given set. We focus on ROI-level classification in this paper.

		Benign	Atypia	In Situ	Invasive	Total
Slide	Training Set	30	16	16	13	39
	Validation Set	15	7	8	6	18
	Test Set	16	8	9	6	21
	Total	61	31	33	25	78
ROI	Training Set	226	55	154	102	537
	Validation Set	109	25	56	50	240
	Test Set	105	30	69	49	253
	Total	440	110	279	201	1030

Since the specimens were prepared at different times, they have a high variation in their staining. Thus, we performed stain normalization by matching the histograms of the hematoxylin and eosin channels of each slide to the hematoxylin and eosin histograms of a target slide chosen from the data set.[12] To obtain hematoxylin and eosin histograms, we applied color deconvolution[13] to each slide using a unique stain matrix estimated for that slide. Hematoxylin stain vector estimation was carried out by computing the median of the pixels inside the nucleus mask of the slide in the optical density space, separately for red, green, and blue channels. For eosin, the median was computed over a mask obtained by eliminating the nuclei and high luminosity regions. We estimated the nucleus masks using a pre-trained convolutional network.

Finally, we partitioned the data set into four folds by using ROI-level diagnostic labels by making sure that each fold has slides (and ROIs) corresponding to independent patients. To achieve both slide-level and ROI-level diversity among the folds, we employed a genetic algorithm which rewards splitting the data set into similar numbers of ROIs and slides per class within each fold. Two randomly chosen folds are combined as the training set, whereas validation and test sets are randomly selected among the remaining folds. The ROI-level and slide-level class distributions of the three sets are given in Table 2. Since each slide has a different number of ROIs, that may also have a set of labels different from that of the slide, the resulting data set has a heavy class imbalance.

3. METHODOLOGY

Problem definition. The ultimate goal is to classify a given ROI image of arbitrary size into one of the diagnostic classes. During training, we have access to example ROI images and their ROI-level class annotations. Each training ROI sample is associated with one particular class label; however, not all patches inside an ROI homogeneously belong to the same class. Therefore, the goal is to learn the classification model in a weakly-supervised manner over the noisy patches and ROI-level annotations, without having access to patch-level labels.

Graph construction. We tackle the ROI classification problem as a graph classification problem, where vertices represent patches and edges represent spatial relations across the patches. In this manner, we aim to aggregate information from patches and admit arbitrary-sized ROIs in a principled manner. For this purpose, we construct an ROI graph, by first regularly sampling fixed-sized patches from the ROI, and associating each vertex with the corresponding image patch. Then, we add a binary edge between each pair of patches that are within a pre-defined proximity threshold ϵ. This leads to a sparse binary ROI graph. Example graphs for several ROIs are shown in Figure 1.

Architecture. We propose a graph convolutional network (GCN) for the ROI graph classification problem. The first part of the network employs a ResNet-50[14] convolutional sub-network that extracts a fixed-length (2048-dimensional) feature vector for each vertex (*i.e.,* patch). In order to propagate information across the patches and incorporate local contextual information, we apply two consecutive GCN layers. Here, we incorporate the GCN layer definition introduced by Kipf and Welling.[15] According to this definition, the GCN layer first calculates the weighted sum of the feature vectors of a vertex and the neighboring vertices of that vertex. Here, the GCN aggregation weights are induced by the symmetrically normalized adjacency matrix.[15] Then, a linear

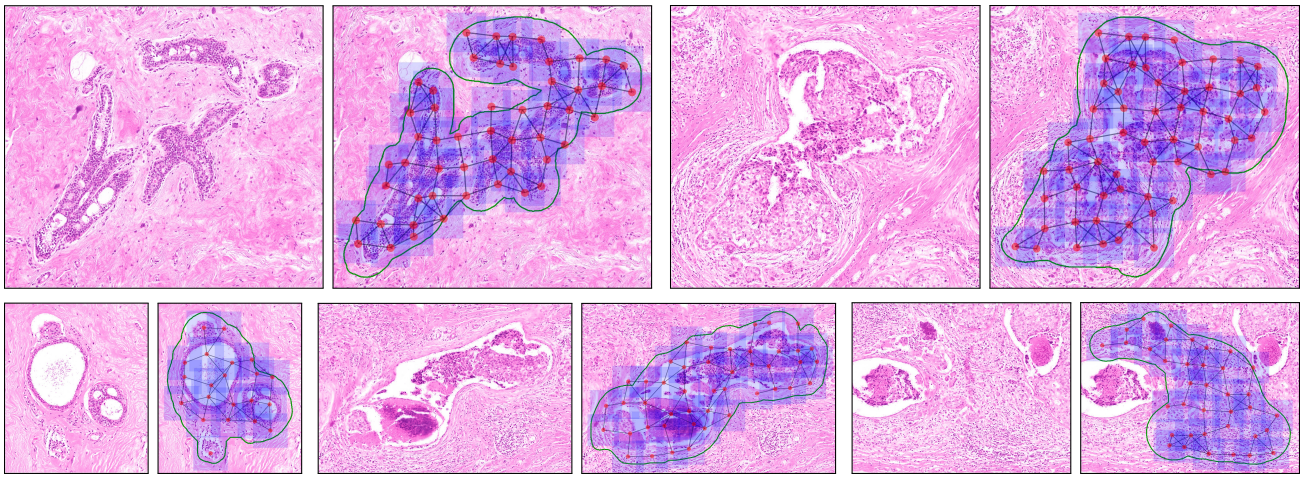

Figure 1. Sample ROIs (left) and the constructed ROI graphs (right). ROI boundaries drawn by the pathologists are shown in green and the sampled patches are displayed in blue. ROI graphs constructed with the proximity threshold ϵ chosen as 200 pixels are overlaid with vertices in red and edges in black.

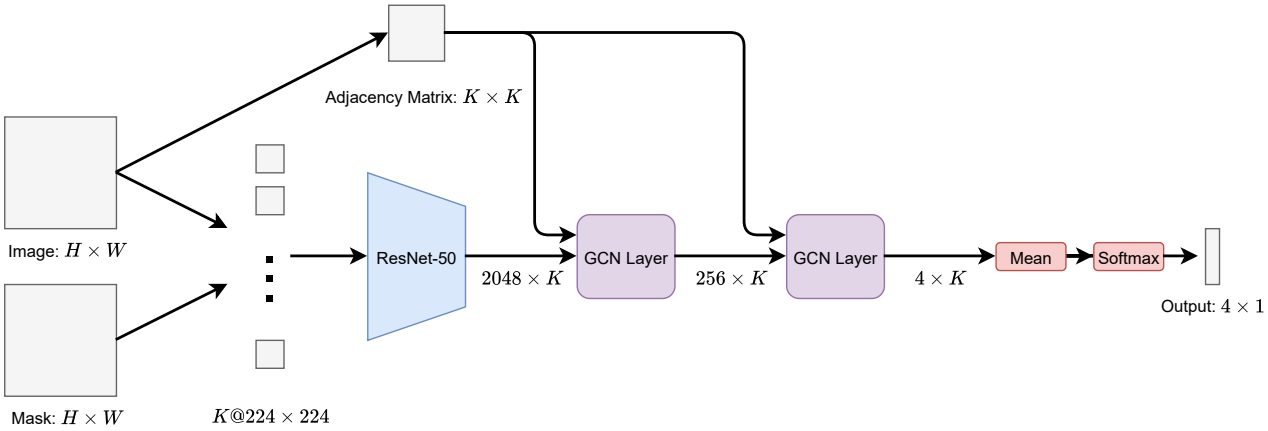

Figure 2. The proposed ROI graph classification architecture. K represents the number of patches (*i.e.,* vertices) inside the ROI, which is represented by the input graph. 4 corresponds to the number of diagnostic classes. Input mask represents the ROI area within the input image.

transformation followed by a nonlinear activation function is applied to these aggregated feature vectors to obtain per-vertex output vectors of the GCN layer. While the first GCN layer transforms each vertex into a 256-dimensional vector, the second one results in 4-dimensional vectors, which can be interpreted as internal per-vertex classification scores. Finally, we apply global average pooling over all resulting vertex vectors to obtain a fixed-length representation for the whole graph, and compute a soft-max to obtain the final ROI classification probabilities. The model is summarized in Figure 2.

Training. Our training pipeline consists of two separate training steps: first, training a patch classifier network to be used as a feature extractor; second, training the GCN layers for ROI classification. For the first step, we label each patch extracted from an ROI with the annotation of that ROI. Using these labels, we fine-tune the ResNet sub-network which is initialized via a classification model pre-trained on the ImageNet data set. In the second step, we train GCN layers with randomly initialized parameters using the features extracted by the ResNet while the ResNet parameters are kept frozen. The training is carried out in a weakly-supervised fashion, purely based on ROI class labels, with no patch-level annotations. For this purpose, we minimize the negative log-likelihood of the true class label of each ROI example through stochastic gradient-descent.

4. EXPERIMENTS

In this section, we provide the training details of the patch-level feature extraction network and the GCN-based contextual classification network, define the baseline methods for empirical comparison, and present our experimental results.

Patch representation learning. For training the ResNet-based patch-level feature extraction model, we sample patches of size 224×224 with an overlap of 74 pixels across consecutive patches at $10\times$ magnification. We over-sample patches from classes with fewer examples to reduce the effect of class imbalance. We apply random horizontal/vertical flips, random rotations of 90 degrees, and random hue jitter for data augmentation. We use the Adam optimizer with a learning rate of 5×10^{-7}, apply weight decay to all ResNet parameters with weight 10^{-3}, and use dropout[16] before the classification layer with 0.7 drop probability. Each batch contains 128 patches. After training the ResNet model in this manner, we keep its parameters frozen during the following GCN training stage.

GCN training. For training the proposed GCN-based contextual ROI classification model, we keep the ResNet model parameters frozen and use pre-extracted patch descriptors in order to simplify training over ROIs with a variety of bounding aspect ratios and sizes. To obtain patch samples from an ROI, we use the same patch sampling and data augmentation techniques as in our aforementioned ResNet training approach with two exceptions: (i) we turn off random hue jittering (just to simplify the training pipeline implementation) and (ii) we sample additional patches by jittering the center coordinates of the original patch samples in both horizontal and vertical directions by random amounts. We obtain the random coordinate jitter values by sampling uniformly from the interval $[-45, 45]$. The size of the first GCN layer is chosen as 256 and the second one as the number of classes. We again use the Adam optimizer with a learning rate of 1×10^{-3}, apply weight decay to all GCN parameters with weight 10, and use dropout in the first GCN layer with 0.6 drop probability. We apply batch normalization after the first GCN layer. Each batch consists of 1024 ROIs chosen from the augmented training set.

A detail that deserves attention is the way to use dropout before a GCN layer: while applying dropout, we jointly process the patches within a single ROI and drop the same dimensions from their feature vectors instead of applying dropout to their descriptors independently. Otherwise (when applied independently), the effect of dropout is diminished due to local feature averaging in the following GCN layer. This is akin to using channel-wise *spatial dropout*[17] in convolutional layers to keep dropout effective without getting diluted due to the correlations across neighboring pixels.

Baselines. We use two versions of the model described in Section 3. The version denoted as $\text{GCN}_{\epsilon=0}$ uses the identity matrix as the adjacency matrix. This corresponds to the GCN layers acting as fully-connected classification layers without any spatial context information. The loss function that is used during the training of the model is still the same as the one described in Section 3. The version denoted as $\text{GCN}_{\epsilon=200}$ uses a threshold of 200 pixels in construction of the adjacency matrix.

We additionally compare the performance of the proposed model with the following methods:

- Base-Penultimate: A patch-level feature representation is extracted directly from the penultimate layer activations of the fine-tuned convolutional network (ResNet).[8] Then, the feature representations of the individual patches inside an ROI are aggregated by average pooling to obtain the feature representation of the ROI. Finally, a multi-layer perceptron (MLP) classifier is trained on the ROI-level feature representations and the labels in the training set.
- Majority-Voting: The fine-tuned convolutional network (ResNet) is used to assign a class label to each patch individually. Then, the class label for an ROI is obtained through majority voting of the class labels of its corresponding patches.
- Learned-Fusion: Similar to Majority-Voting, each patch is assigned a score for each class by the patch classification network (ResNet).[18] Then, a class histogram for each ROI is constructed by summing over the class scores of its corresponding patches. Finally, the extracted histograms are used as feature vectors to be classified through an MLP.

In our comparisons, we use *normalized accuracy*, which is obtained by averaging per-class accuracy scores, to avoid biases towards classes with higher number of samples.

Table 3. Experimental results (in normalized accuracy).

	Normalized Acc. (%)
Majority-Voting	66.13
Learned-Fusion	67.03
Base-Penultimate	71.85
$GCN_{\epsilon=0}$	76.90
$GCN_{\epsilon=200}$	78.56

Table 4. Confusion matrix of $GCN_{\epsilon=200}$.

		Predicted		
		Benign	In Situ	Invasive
Reference	Benign	89	15	1
	In Situ	22	45	2
	Invasive	0	7	42

In our preliminary experiments, we observed that consistently all methods performed poorly on the atypia class. This is likely to be a result of the fact that there are relatively few ROIs belonging to this class in our data set that is acquired from whole slides sampled from routine clinical practice. In addition, distinguishing atypia from in situ or benign cases appears to be a more challenging problem,[19] which is also emphasized in other related work that do not include samples from the atypia class.[3,20] Following these observations, we exclude the ROIs belonging to the atypia class, and conduct our in-detail experiments using the three remaining diagnostic classes: benign, in situ carcinoma, and invasive carcinoma.

Results. The experimental results are summarized in Tables 3 and 4. These results show that the proposed model using the graph convolutional layers in the non-degenerate graph setting ($GCN_{\epsilon=200}$) achieves a considerably higher accuracy than the no-edge case ($GCN_{\epsilon=0}$), with an improvement of 1.5 points. This result suggests that the proposed GCN network effectively leverages the local contextual relations across the patches. We also observe that it outperforms the Majority-Voting, Learned-Fusion, and Base-Penultimate baselines, which suggests that our weakly supervised learning scheme is an effective approach for training with ROI-level labels.

Figure 3 shows example patch classification results for the ResNet-based patch classifier (top row), $GCN_{\epsilon=0}$ model (middle row), and $GCN_{\epsilon=200}$ model (bottom row). For the GCN models, the patch classification results are obtained by using the final unnormalized scores computed before the global average pooling layer. These results shown in the figure highlight that the final $GCN_{\epsilon=200}$ model yields significantly fewer local misclassifications and spatially smoother predictions, compared to both the patch classifier and the $GCN_{\epsilon=0}$. We note that both GCN models are trained to minimize the ROI classification loss, not per-patch classification losses. Therefore, overall, these results also support the hypothesis that the proposed GCN model can learn to leverage the spatial context information encoded in the graph structure.

Finally, we re-look into the problem of the atypia class. Although we conduct our main experiments without considering the ROIs annotated as atypia, in our preliminary experiments on an earlier data set[8] which has 437 ROIs obtained from 240 slides belonging to a similar set of 4 diagnostic classes including atypia, we observe that the proposed model achieves an accuracy of 69.9%. This result is noticeably better than that of the state-of-the-art method of Mercan et al.,[8] which yields an accuracy of 68.0%.* This observation supports that the poor performance in atypia classification in the current data set is indeed a data set problem, and can potentially be addressed by collecting more atypical case examples and increasing the diversity in the data set.

5. CONCLUSIONS

In this paper, we tackle the problem of classifying ROIs of arbitrary shape and size in breast histopathology images. We observe that the mainstream approach of first classifying individual patches and then combining the patch classification results fails to leverage the rich spatial context in complex tissue structures. To address this

*Mercan et al.[8] reports 66.8% unnormalized accuracy. We have obtained 68.0% by re-implementing the same method and computing normalized accuracy.

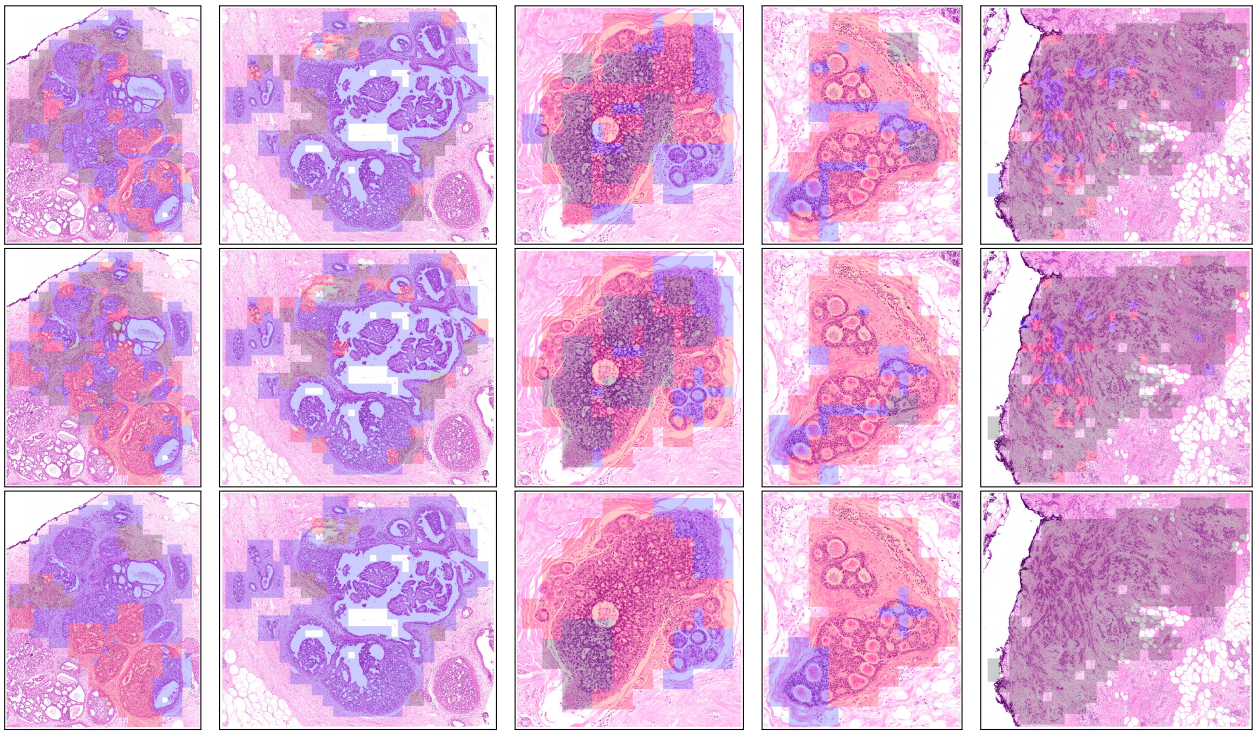

Figure 3. Predictions of the individual patches for sample test ROIs (columns), obtained from the fine-tuned patch classifier network (top), $GCN_{\epsilon=0}$ (middle), and $GCN_{\epsilon=200}$ (bottom). Patches predicted as benign are shown in blue, in situ carcinoma in red, and invasive carcinoma in black. Predictions for the individual patches are obtained just before the global mean pooling layer for the GCN models. The first two ROIs from the left are diagnosed as benign, the next two as in situ, and the last as invasive. The local averaging effect of $GCN_{\epsilon=200}$ is visible at the bottom-most row in the form of smoother patch prediction distributions.

limitation without resizing ROIs into predefined fixed sizes, we propose a graph-based ROI representation and a GCN-based architecture that operate on these graph structures. Our experimental results indicate significant improvements over a number of strong baselines, and suggest that the proposed approach is capable of leveraging the spatial context information in ROIs.

ACKNOWLEDGMENTS

This work was supported in part by the Scientific and Technological Research Council of Turkey (grant 117E172).

REFERENCES

[1] Litjens, G., et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis* **42**, 60–88 (December 2017).

[2] Spanhol, F. A., Oliveira, L. S., Petitjean, C., and Heutte, L., "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering* **63**(7), 1455–1462 (2016).

[3] Aresta, G., et al., "BACH: Grand challenge on breast cancer histology images," *Medical Image Analysis* **56**, 122–139 (2019).

[4] Araújo, T., et al., "Classification of breast cancer histology images using convolutional neural networks," *PloS ONE* **12**(6) (2017).

[5] Roy, K., Banik, D., Bhattacharjee, D., and Nasipuri, M., "Patch-based system for classification of breast histology images using deep learning," *Computerized Medical Imaging and Graphics* **71**, 90–103 (2019).

[6] Das, K., Karri, S. P. K., Roy, A. G., Chatterjee, J., and Sheet, D., "Classifying histopathology whole-slides using fusion of decisions from deep convolutional network on a collection of random multi-views at multi-magnification," in [*IEEE International Symposium on Biomedical Imaging*], 1024–1027 (2017).

[7] Feng, Y., Zhang, L., and Mo, J., "Deep manifold preserving autoencoder for classifying breast cancer histopathological images," *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2019).

[8] Mercan, C., Aksoy, S., Mercan, E., Shapiro, L. G., Weaver, D. L., and Elmore, J. G., "From patch-level to ROI-level deep feature representations for breast histopathology classification," in [*SPIE Medical Imaging Symposium, Digital Pathology Conference*], (2019).

[9] Yan, S., Xiong, Y., and Lin, D., "Spatial temporal graph convolutional networks for skeleton-based action recognition," in [*Thirty-second AAAI Conference on Artificial Intelligence*], (2018).

[10] de Amorim, C. C., Macêdo, D., and Zanchettin, C., "Spatial-temporal graph convolutional networks for sign language recognition," in [*International Conference on Artificial Neural Networks*], 646–657 (2019).

[11] Zhou, Y., Graham, S., Alemi Koohbanani, N., Shaban, M., Heng, P.-A., and Rajpoot, N., "CGC-Net: Cell graph convolutional network for grading of colorectal cancer histology images," in [*IEEE International Conference on Computer Vision Workshops*], (2019).

[12] Basavanhally, A. and Madabhushi, A., "EM-based segmentation-driven color standardization of digitized histopathology," in [*SPIE Medical Imaging Symposium, Digital Pathology Conference*], **8676** (2013).

[13] Ruifrok, A. and Johnston, D., "Quantification of histochemical staining by color deconvolution," *Analytical and Quantitative Cytology and Histology* **23**(4), 291–299 (2001).

[14] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in [*IEEE Conference on Computer Vision and Pattern Recognition*], 770–778 (2016).

[15] Kipf, T. N. and Welling, M., "Semi-supervised classification with graph convolutional networks," in [*5th International Conference on Learning Representations*], (2017).

[16] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R., "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research* **15**(1), 1929–1958 (2014).

[17] Tompson, J., Goroshin, R., Jain, A., LeCun, Y., and Bregler, C., "Efficient object localization using convolutional networks," in [*IEEE Conference on Computer Vision and Pattern Recognition*], 648–656 (2015).

[18] Hou, L., Samaras, D., Kurc, T. M., Gao, Y., Davis, J. E., and Saltz, J. H., "Patch-based convolutional neural network for whole slide tissue image classification," in [*IEEE Conference on Computer Vision and Pattern Recognition*], 2424–2433 (2016).

[19] Elmore, J. G., Longton, G. M., Pepe, M. S., Carney, P. A., Nelson, H. D., Allison, K. H., Geller, B. M., Onega, T., Tosteson, A. N. A., Mercan, E., Shapiro, L. G., Brunye, T. T., Morgan, T. R., and Weaver, D. L., "A randomized study comparing digital imaging to traditional glass slide microscopy for breast biopsy and cancer diagnosis," *Journal of Pathology Informatics* **8**(1), 1–12 (2017).

[20] Bejnordi, B. E., Balkenhol, M., Litjens, G., Holland, R., Bult, P., Karssemeijer, N., and Van Der Laak, J. A., "Automated detection of DCIS in whole-slide H&E stained breast histopathology images," *IEEE Transactions on Medical Imaging* **35**(9), 2141–2150 (2016).