# Finding a Needle in the Haystack: Attention-Based Classification of High Resolution Microscopy Images

Naofumi Tomita[1], Behnaz Abdollahi[1], Jason Wei[1,2], Bing Ren[3], Arief Suriawinata[3], Saeed Hassanpour[†1,2,4]
[1]Dept. of Biomedical Data Science, Dartmouth College
[2]Dept. of Computer Science, Dartmouth College
[3]Dept. of Pathology and Laboratory Medicine, Dartmouth-Hitchcock Medical Center
[4]Dept. of Epidemiology, Dartmouth College
[1,2,3]{naofumi.tomita, behnaz.abdollahi, jason.20, saeed.hassanpour}@dartmouth.edu
[4]{bing.ren, arief.a.suriawinata}@hitchcock.org

## Abstract

*Deep learning for classification of microscopy images is challenging because whole-slide images are high resolution. Due to the large size of these images, they cannot be transferred into GPU memory, so there are currently no end-to-end deep learning architectures for their analysis. Existing work has used a sliding window for crop classification, followed by a heuristic to determine the label for the whole slide. This pipeline is not efficient or robust, however, because crops are analyzed independently of their neighbors and the decisive features for classifying a whole slide are only found in a few regions of interest. In this paper, we present an attention-based model for classification of high resolution microscopy images. Our model dynamically finds regions of interest from a wide-view, then identifies characteristic patterns in those regions for whole-slide classification. This approach is analogous to how pathologists examine slides under the microscope and is the first to generalize the attention mechanism to high resolution images. Furthermore, our model does not require bounding box annotations for the regions of interest and is trainable end-to-end with flexible input. We evaluated our model on a microscopy dataset of Barrett's Esophagus images, and the results showed that our approach outperforms the current state-of-the-art sliding window method by a large margin.*

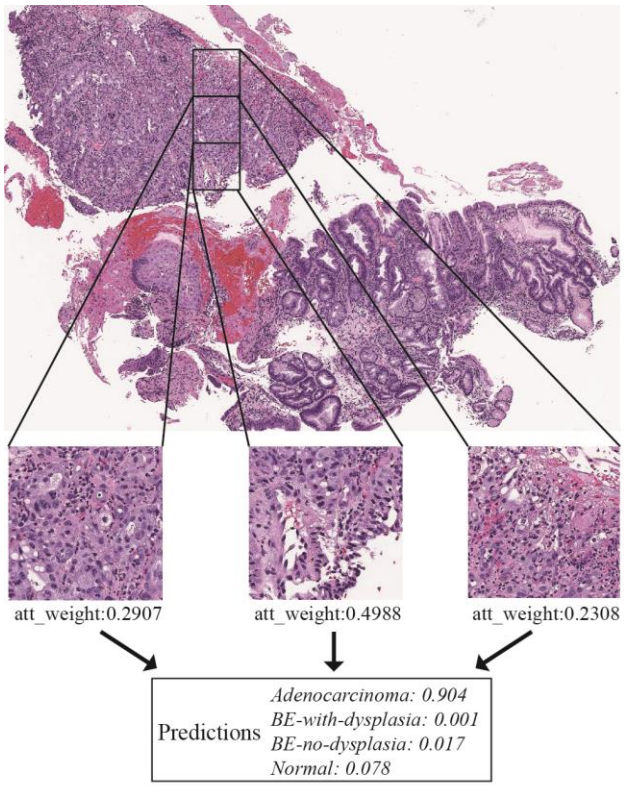


Figure 1. Our attention-based model dynamically finds regions of interest for closer inspection. Predictions are made based on weighted features from each tile. The size of this microscopy image is 4,428 × 6,396 pixels.

## 1. Introduction

In the field of pathology, tissue slides are scanned as high resolution images, which can have sizes up to 10,000 × 10,000 pixels. This high resolution is necessary because each whole slide contains thousands of cells, for which the cellular structures must be visible in order to identify regions of the tissue that indicate disease (lesions). However, the size of lesions is often relatively small, typically around 100 × 100 pixels, as most of the cells in a given slide are normal. Therefore, the decisive regions of interest containing lesions usually comprise much less than one percent of the tissue area. Even for trained pathologists, localizing these lesions for the classification of the whole slide is time consuming and often inconsistent.

In recent years, deep learning has made considerable

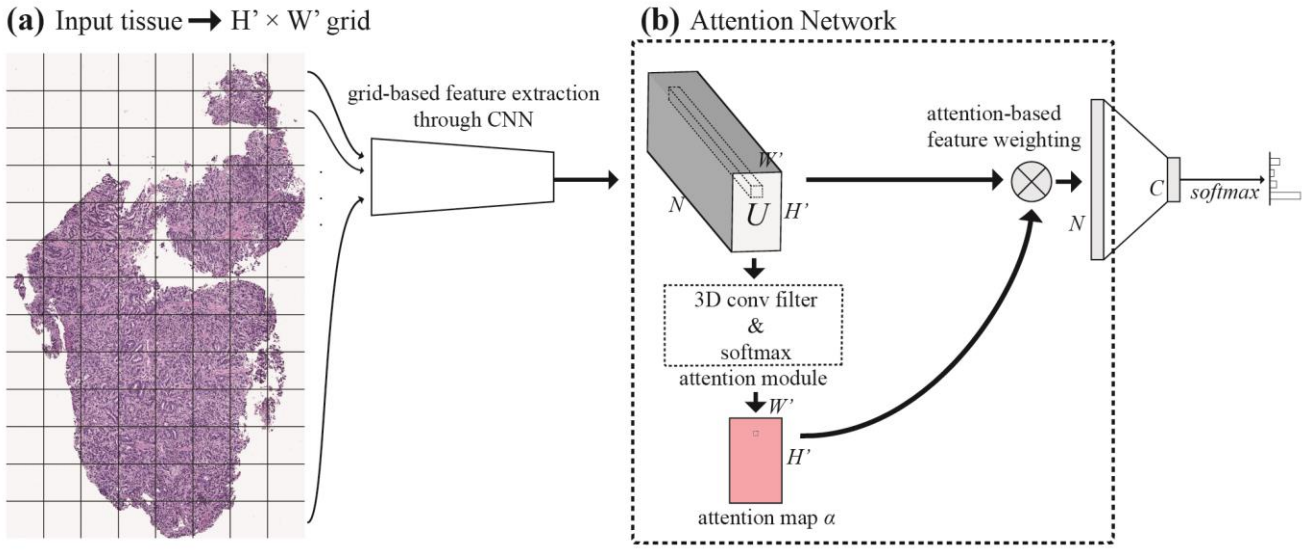


Figure 2. Overview of our attention-based model. (a) An input image $\boldsymbol{x}$ is divided into $\boldsymbol{H' \times W'}$ grid cells (dividing lines are shown only for visualization). (b) Learnable 3D convolutional filters of size $\boldsymbol{N \times d \times d}$ are applied on the grid-based feature map tensor $\boldsymbol{U}$ to generate an attention map $\boldsymbol{\alpha}$, which operates as the weights for an affine combination of $\boldsymbol{U}$.

advances in classification of microscopy images. The most common approach in this domain involves a sliding window for crop classification, followed by statistical methods of aggregation for whole-slide inference [4, 12, 15, 16, 24, 29]. In this approach, pathologists annotate bounding boxes on whole slides in order to train a classifier on small crops, typically of sizes in the range of $200 \times 200$ pixels to $500 \times 500$ pixels. For evaluating a whole slide, this crop classifier is applied to extracted windows from the image, and then a heuristic, often developed in conjunction with a domain-expert pathologist, is used to determine how the distribution of crop classification scores translates into a whole-slide diagnosis.

However, there are many limitations to this sliding window approach. The first is that since crop classifiers are needed, all images in the training set must be annotated by pathologists with bounding boxes around each region of interest. In addition, developing a heuristic for aggregating crop classifications often requires pathologist insight. This is possible when engineers have easy access to medical professionals, but is not scalable, the heuristics used are dependent on the nature of the classification task and therefore unique. Finally, in the sliding window approach, crops are classified independently of their neighbors and whole-slide classification does not consider the correlations between neighboring windows.

In this paper, we present a model that uses an attention-based mechanism to classify microscopy images. Figure 1 shows the overview of this attention-based classification model. Our approach has the following contributions:

- Our model dynamically identifies regions of interest in a high resolution image and makes a whole-slide classification based on analyzing only these selected regions. This is analogous to how pathologists examine slides under the microscope.
- Our model is trainable end-to-end with only whole-slide labels. All components of our model are optimized through backpropagation. Unlike the current sliding window approach, our model does not need bounding box annotations for regions of interest or pathologist insight for heuristic development.
- Our model is flexible with regard to input size for images. Inspired by fully convolutional network philosophy [19], our grid-based attention module uses a 3D convolution operation that does not require a fixed size input grid. The input size can be any rectangular shape that fits in GPU memory.

Our model is also the first to generalize the attention mechanism to high resolution image classification. Figure 2 and Figure 3 summarize the steps of our proposed model.

## 2. Related Work

**Attention mechanisms.** Our work is inspired by attention models applied to regular image analysis tasks, especially image captioning [1, 2]. Attention mechanisms are described as a part of the prediction module that sequentially selects subsets of input to be processed [2].

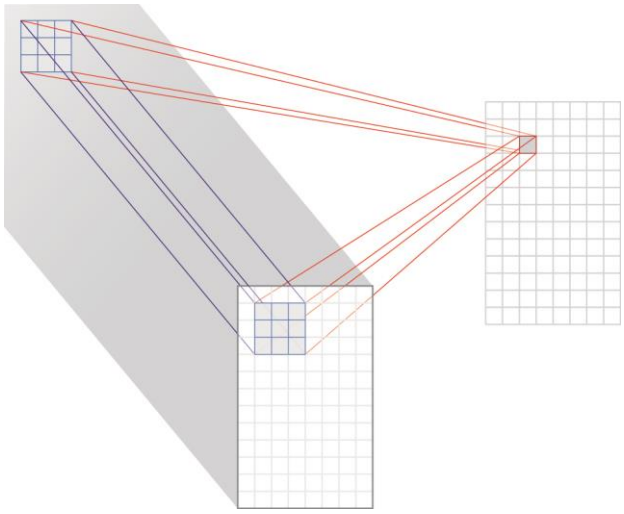

Figure 3. Our grid-based attention mechanism uses a 3D convolution. The significance of each location (right) is determined based on its own features and those of its surrounding crops (the blue box on left). In this figure, the application of a 3D convolutional filter of size $\mathbf{512 \times 3 \times 3}$ is depicted by red lines.

Although this definition is not applicable to non-sequential tasks, the essence of attention mechanisms can be reduced to the ability of networks to generate a dynamic representation of features through weighting them in response to the geometric and holistic context of input. Recent advancement of soft attention enabled end-to-end training on convolutional neural network (CNN) models [1, 6, 14, 28]. Spatial transformer networks capture high-level information from inputs to derive affine transformation parameters, which are subsequently applied to spatial invariant input for a CNN [14]. For semantic segmentation tasks, the attention mechanism is applied to learn multi-scale features [1]. Residual attention networks use soft attention masks to extract features in different granularities [28]. To analyze images in detail, a top-down recurrent attention CNN has been proposed [6]. Our work is based on the soft attention mechanism, but designed for classification of high resolution images that are not typically encountered in the field of computer vision.

**Attention in medical image analysis.** There have been several applications of the attention mechanism in the medical domain, such as using soft attention to generate masks around lesion areas on CT images [8] and employing recurrent attention models fused with reinforcement learning to locate lung nodules [23] or enlarged hearts [30] in chest radiography images. In pathology, recorded navigation of pathologists has been used as attention maps to detect carcinoma [3]. A soft attention approach in [8] deploys two parallel networks for the classification of thorax disease. We draw inspiration from this work but directly reuse extracted features in a single attention network.

## 3. Model

Our proposed approach has two phases. The first phase is grid-based feature extraction from the whole image, where we look at each grid cell in the whole slide to generate a feature map. In the second phase, we apply our proposed attention strategy on the extracted features for whole-slide classification. Notably, the feature extractor is jointly optimized across all the tiles along with the attention module in an end-to-end fashion.

### 3.1. Grid-based Feature Extraction

To extract features on the whole image through a CNN, we divide every image into smaller tiles with no overlap. Features are extracted from each tile and reformatted to a single grid-based set of features. We generate this feature map in the following fashion: let $x$ denote an input image of shape $3 \times H \times W$, where $3$, $H$, and $W$ are the RGB channels, height, and width of the image, respectively. Through feature extraction, we obtain a feature map tensor $U$ of shape $N \times H' \times W'$, where $N$ is the number of extracted features, $H'$ is the number of rows, and $W'$ is the number of columns of non-overlapping tiles on the image. Specifically, $H' \coloneqq \lfloor H/h \rfloor$ and $W' \coloneqq \lfloor W/w \rfloor$, where $h$ and $w$ are the height and width of each tile.

In terms of CNN architecture, we use the residual neural network (ResNet) architecture [10], one of the state-of-the-art CNN models with high performance on the ImageNet Large Scale Visual Recognition Competition (ILSVRC) as well as many medical image classification tasks [2, 31]. Among several variants of ResNet models, we choose the pre-activation ResNet-18 model [11]. This model achieves a good trade-off between performance and GPU memory usage, which is vital for processing high resolution image data. By removing the final fully-connected layer before the global pooling layer, the network produces a tensor of size $512 \times H' \times W'$ as output for an input image. We extend this model by replacing all 2D convolutions with 3D convolutional filters of shape $1 \times 3 \times 3$ in order to implement mini-batch training for image samples. Consequently, input tensors are $T \times (H' \cdot W') \times 3 \times h \times w$, where $T$ is mini-batch size.

### 3.2. Attention-based Classification

After feature extraction, attention modules are applied to the feature map, with their weights determining the importance of each tile. Then, we compute a feature vector and optimize against labels of each image in a feedforward

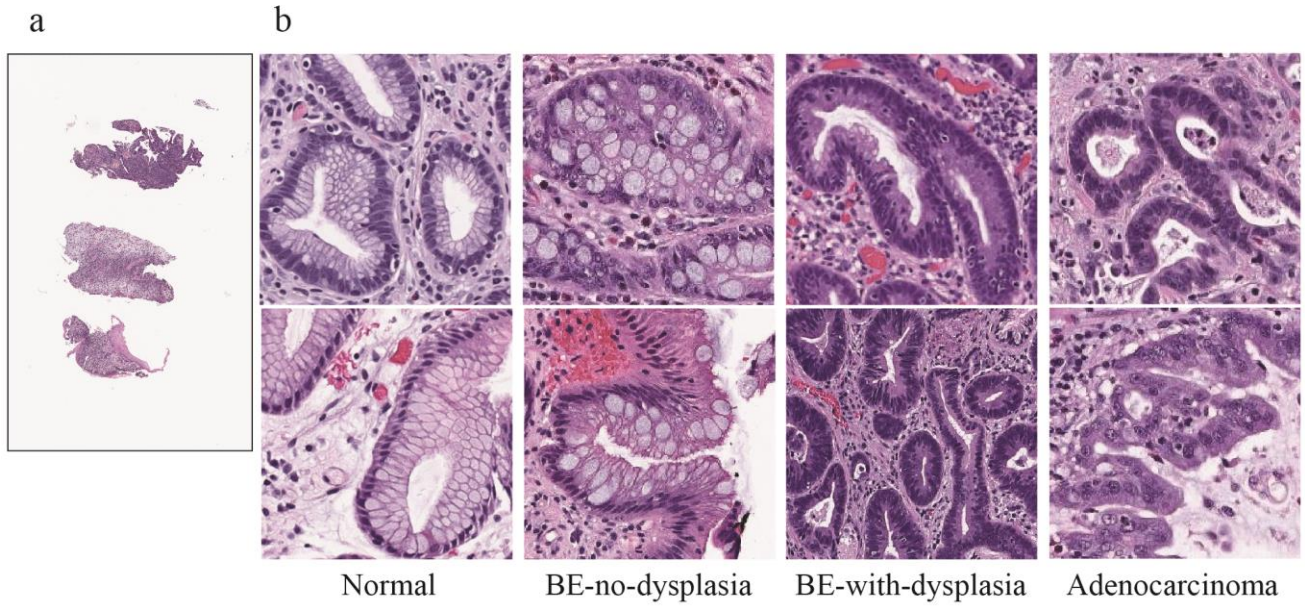


Figure 4. (a) A typical whole-slide image in our dataset. This particular slide contains three separate tissues and is of size 9,440 × 15,340 pixels. (b) Samples from each class of Barrett's Esophagus. A clinical description of classification system can be found in Appendix A.

neural network, allowing for classification of the entire whole-slide image. For our attention mechanism, we apply 3D convolutional filters of size N × d × d , where $N$ is the kernel's depth and $d$ denotes the height and width of the kernels. Without loss of generality, we can consider a case with one filter and one corresponding attention map. Applying a 3D filter of size $N \times d \times d$ to a feature map $U$ produces a raw attention output of $V \in \mathbb{R}^{H' \times W'}$. Once $V$ is computed, an attention map is calculated by:

$$\boldsymbol{\sigma}(\boldsymbol{V})_{i,j} = \frac{e^{V_{i,j}}}{\sum_{h=1}^{H'} \sum_{w=1}^{W'} e^{V_{h,w}}} \quad (1)$$

where $i$ and $j$ are row and column indices of the resulting attention map $\alpha$. By treating the attention map $\alpha$ as feature weights, the components $z_n$ of the final feature vector $z$ are computed by:

$$\boldsymbol{z_n} = \sum_{h=1}^{H'} \sum_{w=1}^{W'} \boldsymbol{\sigma}(\boldsymbol{V})_{h,w} \cdot \boldsymbol{U}_{n,h,w} \quad (2)$$

The feature vector $z$ is subsequently used for whole-slide classification through fully connected layers and a non-linear activation function.

Moreover, the use of multiple attention modules in our framework can potentially capture more local patterns for classification, increasing the capacity and robustness of the network, especially for medical images of high resolution. As such, we simultaneously apply $m$ 3D filters that generate $m$ attention maps and individually populate $m$ feature vectors. All feature vectors are concatenated to form a single vector, which is fed to the fully connected classifier.

In the end-to-end training pipeline, the cross-entropy loss over all classes is computed on classification predictions. The loss is backpropagated to optimize all parameters in the network without any specific adjustment for attention modules. Our model does not need bounding box annotations around regions of interest, and all optimization is done with respect to only the labels at the whole-slide level.

## 4. Experiments

To evaluate our attention-based classification model for high resolution microscopy images, we applied our method to a microscopy dataset of Barrett's Esophagus (BE) images, which are slides of tissues surgically removed from patients at risk of esophageal cancer. We compared the results of our proposed model's performance to those generated by the state-of-the-art sliding window method [5, 16, 18]. We found that our model outperforms this sliding window model by a large margin.

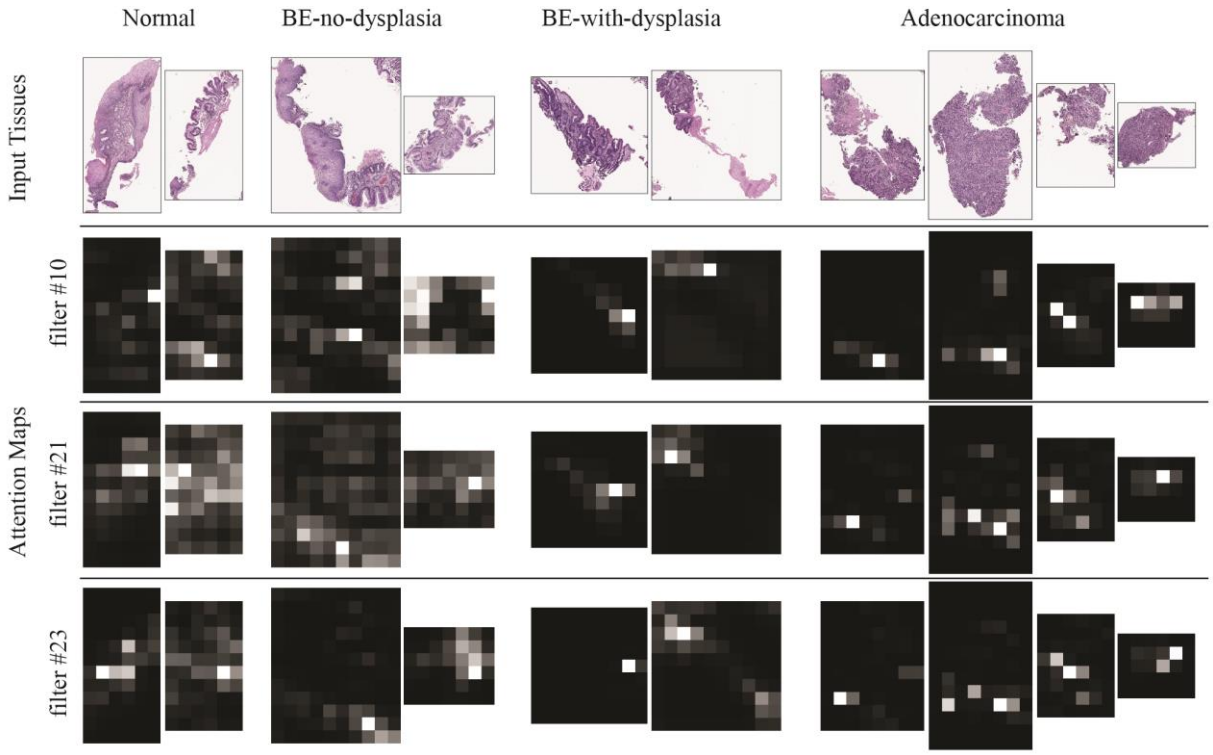


Figure 5. Examples of attention maps generated by different attention modules (filters) which are optimized for attending to the features of the Adenocarcinoma class. Top row shows whole-slide sub-images from the test set. The second to fourth rows show attention maps of the selected attention modules for input images from different ground truth classes. Higher attention weight is denoted by white color and lower is denoted by black color. For visualization purposes, each map is normalized so its maximum value is 1. The attended regions for the Adenocarcinoma class images are verified to be correct by two pathologists. In contrast, these attention modules are inattentive to lower risk class images.

### 4.1. Dataset

For this experiment, whole-slide images were collected from patients who underwent endoscopic gastric mucosal biopsy since 2017 at our academic medical center. Leica Aperio scanners were used to digitize H&E-stained whole-slide images at 20 × magnification. We had a total of 180 whole-slide images. 116 were used as the development set and 64 were used as the test set. 20% of the development set were reserved for validation.

In order to determine labels for whole-slide images and to train the sliding window method as our baseline, bounding boxes around lesions in these images were annotated by two pathologists from our academic institution. We considered these labels as reference standard, as any disgreements in annotation were resolved through further discussion among annotators and consultation with a senior domain-expert pathologist. These bounding boxes were not used in training our attention-based model.

For preprocessing, we removed white background on the slides and extracted only regions of the images that contain tissue. Figure 4a shows a typical whole-slide image from our dataset. These images can cover multiple pieces of tissue, so we separated them into large sub-images with average size of $5{,}131 \times 5{,}875$ pixels, each only covering a single piece of tissue. Every sub-image was given an overall label based on the labels of its lesions. If multiple lesions with different classes were present, we used the class with the highest risk as the corresponding label, as that lesion would have the highest impact clinically. If no abnormal lesions were found in a sub-image, it was assigned to the normal class. After this preprocessing step, each sub-image was assigned to one of our four classes: Normal, BE-no-dysplasia, BE-with-dysplasia, and Adenocarcinoma (Figure 4b). Our dataset included 256 sub-images after preprocessing. To avoid possible data leakage, extracted tissues from one whole-slide image were all placed into the same set of images when the development and test set were split. Table 1 summarizes the test set.

| Diagnosis | Number (%) |
|---|---|
| Normal | 58 (47.2%) |
| BE-no-dysplasia | 30 (24.4%) |
| BE-with-dysplasia | 14 (11.4%) |
| Adenocarcinoma | 21 (17.1%) |

Table 1. Class distribution of Barrett's Esophagus images in our test set.

| Ground Truth | | Sliding Window [5, 16, 18] | Attention Model |
|---|---|---|---|
| Normal | Accuracy | 0.63 | **0.70** |
| | Recall | 0.62 | **0.69** |
| | Precision | 0.60 | **0.68** |
| | F1 Score | 0.61 | **0.68** |
| BE-no-dysplasia | Accuracy | 0.78 | **0.82** |
| | Recall | 0.43 | **0.77** |
| | Precision | **0.87** | 0.68 |
| | F1 Score | 0.58 | **0.72** |
| BE-with-dysplasia | Accuracy | 0.68 | **0.83** |
| | Recall | **0.36** | 0.21 |
| | Precision | 0.16 | **0.50** |
| | F1 Score | 0.22 | **0.30** |
| Adenocarcinoma | Accuracy | 0.87 | **0.88** |
| | Recall | 0.52 | **0.71** |
| | Precision | **0.65** | 0.63 |
| | F1 Score | 0.58 | **0.67** |
| Mean | Accuracy | 0.74 | **0.81** |
| | Recall | 0.48 | **0.60** |
| | Precision | 0.57 | **0.62** |
| | F1 Score | 0.50 | **0.63** |

Table 2. Classification results for our test set on Barrett's Esophagus. We assessed the model's performance in accuracy, recall, precision, and F1 score. Results are rounded to two decimal places. Our method outperforms the sliding window baseline in both accuracy and F1 score for all classes.

### 4.2. Sliding Window Baseline

In order to compare our model to previous methods for high resolution image analysis, we implemented the current state-of-the-art sliding window method as described in [5, 16, 18]. In this method, we used our annotated bounding box labels to generate small crops of size $224 \times 224$ pixels for training a crop classifier. For preprocessing, we normalized the color channels and performed standard data augmentation including color jittering, random flips, and rotations.

For training, we initialized ResNet-18 with the He initialization [9]. We optimized over the cross-entropy loss function for 100 epochs, employing standard weight regularization techniques and learning rate decay. We trained our crop classifier to predict the class of any given window in a whole-slide image. For whole-slide inference, we performed a grid search over our validation set to find optimal thresholds for filtering noise. Then, we consulted two separate pathologists to develop heuristics for aggregating crop predictions. We chose the thresholds and heuristic from the approach that performed the best on the validation set and applied that to the whole-slide images in the test set. Performance metrics for this sliding window approach are shown in Table 2.

### 4.3. Attention Model

We implemented our attention model as described in Section 3. Given the size of features extracted from ResNet-18 model, we used $512 \times 3 \times 3$ 3D convolutional filters in the attention module, with the implicit zero-padding of $(0, 1, 1)$ for depth, height, and width dimensions. We employed 64 of these filters to increase the robustness of the attention module, as patterns in the feature space are likely too complex to be recognized and attended by a single filter. To avoid overfitting and encourage each filter to capture different patterns, we regularized the attention module by applying dropout [27] with $p = 0.5$ after concatenating all the feature vectors $z$. We initialized the entire network with the He initialization for convolutional filters [9], unit weight and zero-bias for batch normalizations [13], and the Glorot initialization for fully connected layers [7]. We emphasize that only the cross-entropy loss against class labels is used in training. Other information such as the location of bounding boxes was not given to the network as a guidance to optimal attention maps. Our model identified such regions of interest automatically.

We first initialized our feature extraction network with weights pretrained on the ImageNet dataset [17]. Input of the network was an extracted grid cell of $492 \times 492$ pixels and resized to $224 \times 224$ pixels. We normalized the input values by the mean and standard deviation of pixel values computed over all tissues. The last fully connected layer of the network was removed, and all residual blocks except for the last one were frozen, serving as a regularization mechanism.

We trained the entire network on large tissue images extracted from whole slides. For data augmentation, we applied random rotation and random scaling with a scaling factor between 0.8 and 1.2 on the fly during training. We used the Adam optimizer with an initial learning rate of 1e-3, decaying by 0.95 after each epoch, and reset the learning rate to 1e-4 every 50 epochs in a total of 200 epochs, similar to the cyclical learning rate [20, 26]. We set the mini batch size to 2 to maximize the utilization of memory on our Nvidia Titan Xp GPU. The model was implemented in PyTorch [22].

### 4.4. Results

**Our model outperforms previous methodology.** We performed both a quantitative and a qualitative evaluation of our model. As a reference baseline, we referred to results from using the sliding-window method [16] for this classification task, trained on the same data split but with annotated bounding boxes. For quantitative evaluation, we used four standard metrics for classification: accuracy, recall, precision, and F1 score. Our classification results on the test set are summarized on Table 2. Compared to the baseline, our model achieved better accuracy and F1 score

in all classes. Especially for F1 score, which is the harmonic mean of precision and recall, our model outperformed the baseline approach by at least 8% for each class. Quantitative analysis showed exemplary performance of our model on the Normal, BE-no-dysplasia, and Adenocarcinoma classes. However, both our attention model and the baseline model did not perform well on identifying images of the class BE-with-dysplasia. This is possibly because BE-with-dysplasia was the least frequent class in our dataset, comprising only 11% of images. Of note, our model is also the first to automate classification of tissue in Barrett's Esophagus using histopathology slides.

**Qualitative analysis.** We visualized the generated 64 attention maps for all the testing images to verify the attention mechanism in our model. We present characteristic examples for the Adenocarcinoma class on Figure 5. The distributions of the attention module highlighted across different classes indicate that each module looks for specific features in the Adenocarcinoma class. Furthermore, multiple attention modules complement each other to make a robust classification decision. For images without the target features, the response is low over all regions (the first and second columns). For the third column, we observe that the attention map is focused on specific regions, which is reasonable from a clinical perspective, in which BE-with-dysplasia progresses to Adenocarcinoma as neoplastic epithelia begin to invade the muscularis mucosae [21].

### 4.5. Limitations

Our method has limitations. In terms of our dataset, one limitation is that all experiments were conducted on slides collected from a single medical center and scanned with the same equipment. Another is that our dataset is still relatively small in comparison to conventional datasets in deep learning; in particular, the number of slides of BE-with-dysplasia was small, resulting in lower performance for that class. In order to evaluate the robustness and generalizability of our approach, further verification with different classification tasks and datasets from other institutions is required and will be pursued.

Furthermore, even with our method that is built to analyze entire tissue regions, current GPUs do not have enough memory capacity to process some very large images. For such slides, we divided the tissue area into manageable sub-tissue images and relabeled them. In our experiments, a GPU with 12 GB of memory could process 84.0% of the slides in our dataset. We speculate that a GPU with 48 GB of memory would process 98.6% of the slides in the dataset. Alternatively, the feature extractor, which is the largest source of memory consumption in our approach, could be optimized to address this issue. The ResNet-18 architecture used in our model achieved high performance with a relatively low number of parameters. However, we believe that there is still room for the further reduction of parameters while maintaining high performance. Our model is a novel approach for end-to-end classification of microscopy images, paving the road for future work in using deep learning for the analysis of high resolution images.

## 5. Conclusion

We presented an attention-based model for classification of high resolution microscopy images. Analogous to how pathologists examine slides under the microscope, our model finds regions of interest and examines their features for whole-slide classification. We showed that our model outperforms the current sliding window method on a dataset for Barrett's Esophagus. Previous methodology for analyzing microscopy images is limited by manual annotation and access to medical expertise. Our model, on the other hand, is trained end-to-end with only labels at the whole-slide level, removing the high cost of data annotation and opening the door for deep learning to solve more classification problems in pathology.

## Acknowledgment

This research was supported in part by a National Institutes of Health grant, P20GM104416. The authors would like to thank Lamar Moss and Maksim Bolonkin for their help with this manuscript.

## Appendix A

This project used categories of esophageal cancer as defined by the Vienna classification system [25]. The use of human subject data in this project is approved by our Institutional Review Board (IRB) and the conducted research in this paper is in compliance with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects.

**Normal**: includes normal stratified squamous epithelium, normal squamous and columnar junction epithelium, and normal columnar epithelium.

**BE-no-dysplasia**: includes Barrett's Esophagus negative for dysplasia or indefinite for dysplasia. Barrett's Esophagus is defined by columnar epithelium with goblet cells (intestinal metaplasia) and preservation of orderly glandular architecture of the columnar epithelium with surface maturation. Indefinite for dysplasia denotes that the lesion is suggestive of but not diagnostic of dysplasia, such as significant atypia with or without surface maturation in the context of inflammation, ulceration, or regenerative changes.

**BE-with-dysplasia**: includes noninvasive low-grade neoplasia (low-grade dysplasia) and noninvasive high-grade neoplasia (high-grade dysplasia). Columnar epithelium with low-grade dysplasia is characterized by nuclear pseudostratification, mild to moderate nuclear hyperchromasia and irregularity, and the cytologic atypia extending to the surface epithelium. High-grade dysplasia demonstrates marked cytologic atypia including loss of polarity, severe nuclear enlargement and hyperchromasia, numerous mitotic figures, and architectural abnormalities such as lateral budding, branching, villous formation, as well as variation of the size and shape of crypts.

**Adenocarcinoma**: includes invasive carcinoma (intramucosal carcinoma and submucosal carcinoma and beyond) and suspicious for invasive carcinoma. Cases of high-grade dysplasia with features suggestive of invasion are classified into this category; and the worrisome features include cribriform/solid growth, ulceration occurring within high-grade dysplasia, dilated dysplastic glands with necrotic debris, large angulated glands, and dysplastic tubules incorporated into overlying squamous epithelium.

## References


[1] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3640-3649.

[2] Y.-A. Chung and W.-H. J. a. p. a. Weng, "Learning Deep Representations of Medical Images using Siamese CNNs with Application to Content-Based Image Retrieval," 2017.

[3] G. Corredor, J. Whitney, V. L. A. Pedroza, A. Madabhushi, and E. R. J. J. o. M. I. Castro, "Training a cell-level classifier for detecting basal-cell carcinoma by combining human visual attention maps with low-level handcrafted features," vol. 4, no. 2, p. 021105, 2017.

[4] E. Cosatto *et al.*, "Automated gastric cancer diagnosis on h&e-stained sections; ltraining a classifier on a large scale with multiple instance machine learning," in *Medical Imaging 2013: Digital Pathology*, 2013, vol. 8676, p. 867605: International Society for Optics and Photonics.

[5] N. Coudray *et al.*, "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," vol. 24, no. 10, p. 1559, 2018.

[6] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *CVPR*, 2017, vol. 2, p. 3.

[7] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249-256.

[8] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. J. a. p. a. Yang, "Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification," 2018.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026-1034.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*, 2016, pp. 630-645: Springer.

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

[13] S. Ioffe and C. J. a. p. a. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015.

[14] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," in *Advances in neural information processing systems*, 2015, pp. 2017-2025.

[15] D. Komura, S. J. C. Ishikawa, and S. B. Journal, "Machine learning methods for histopathological image analysis," vol. 16, pp. 34-42, 2018.

[16] B. Korbar *et al.*, "Looking Under the Hood: Deep Neural Network Visualization to Interpret Whole-Slide Image Analysis Outcomes for Colorectal Polyps," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, 2017, pp. 821-827: IEEE.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097-1105.

[18] Y. Liu *et al.*, "Detecting cancer metastases on gigapixel pathology images," 2017.

[19] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431-3440.

[20] I. Loshchilov and F. J. a. p. a. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," 2016.

[21] R. D. Odze and J. R. Goldblum, *Odze and Goldblum Surgical Pathology of the GI Tract, Liver, Biliary Tract and Pancreas E-Book*. Elsevier Health Sciences, 2014.

[22] A. Paszke, S. Gross, S. Chintala, and G. Chanan, "PyTorch," ed, 2017.

[23] E. Pesce, P.-P. Ypsilantis, S. Withey, R. Bakewell, V. Goh, and G. J. a. p. a. Montana, "Learning to detect chest radiographs containing lung nodules using visual attention networks," 2017.

[24] M. Saha, C. Chakraborty, D. J. C. M. I. Racoceanu, and Graphics, "Efficient deep learning model for mitosis detection using breast histopathology images," vol. 64, pp. 29-40, 2018.

[25] R. Schlemper *et al.*, "The Vienna classification of gastrointestinal epithelial neoplasia," vol. 47, no. 2, pp. 251-255, 2000.

[26] L. N. Smith, "Cyclical learning rates for training neural networks," in *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, 2017, pp. 464-472: IEEE.

[27] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. J. T. J. o. M. L. R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," vol. 15, no. 1, pp. 1929-1958, 2014.
[28] F. Wang *et al.*, "Residual attention network for image classification," 2017.
[29] J. Xu, X. Luo, G. Wang, H. Gilmore, and A. J. N. Madabhushi, "A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images," vol. 191, pp. 214-223, 2016.
[30] P.-P. Ypsilantis and G. J. a. p. a. Montana, "Learning what to look in chest X-rays with a recurrent visual attention model," 2017.
[31] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang, "Mdnet: A semantically and visually interpretable medical image diagnosis network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6428-6436.