JAMA Network Open

Original Investigation | Gastroenterology and Hepatology

Attention-Based Deep Neural Networks for Detection of Cancerous and Precancerous Esophagus Tissue on Histopathological Slides

Naofumi Tomita, MS; Behnaz Abdollahi, PhD; Jason Wei; Bing Ren, MD; Arief Suriawinata, MD; Saeed Hassanpour, PhD

Abstract

IMPORTANCE Deep learning–based methods, such as the sliding window approach for cropped-image classification and heuristic aggregation for whole-slide inference, for analyzing histological patterns in high-resolution microscopy images have shown promising results. These approaches, however, require a laborious annotation process and are fragmented.

OBJECTIVE To evaluate a novel deep learning method that uses tissue-level annotations for high-resolution histological image analysis for Barrett esophagus (BE) and esophageal adenocarcinoma detection.

DESIGN, SETTING, AND PARTICIPANTS This diagnostic study collected deidentified high-resolution histological images (N = 379) for training a new model composed of a convolutional neural network and a grid-based attention network. Histological images of patients who underwent endoscopic esophagus and gastroesophageal junction mucosal biopsy between January 1, 2016, and December 31, 2018, at Dartmouth-Hitchcock Medical Center (Lebanon, New Hampshire) were collected.

MAIN OUTCOMES AND MEASURES The model was evaluated on an independent testing set of 123 histological images with 4 classes: normal, BE-no-dysplasia, BE-with-dysplasia, and adenocarcinoma. Performance of this model was measured and compared with that of the current state-of-the-art sliding window approach using the following standard machine learning metrics: accuracy, recall, precision, and F1 score.

RESULTS Of the independent testing set of 123 histological images, 30 (24.4%) were in the BE-no-dysplasia class, 14 (11.4%) in the BE-with-dysplasia class, 21 (17.1%) in the adenocarcinoma class, and 58 (47.2%) in the normal class. Classification accuracies of the proposed model were 0.85 (95% CI, 0.81-0.90) for the BE-no-dysplasia class, 0.89 (95% CI, 0.84-0.92) for the BE-with-dysplasia class, and 0.88 (95% CI, 0.84-0.92) for the adenocarcinoma class. The proposed model achieved a mean accuracy of 0.83 (95% CI, 0.80-0.86) and marginally outperformed the sliding window approach on the same testing set. The F1 scores of the attention-based model were at least 8% higher for each class compared with the sliding window approach: 0.68 (95% CI, 0.61-0.75) vs 0.61 (95% CI, 0.53-0.68) for the normal class, 0.72 (95% CI, 0.63-0.80) vs 0.58 (95% CI, 0.45-0.69) for the BE-no-dysplasia class, 0.30 (95% CI, 0.11-0.48) vs 0.22 (95% CI, 0.11-0.33) for the BE-with-dysplasia class, and 0.67 (95% CI, 0.54-0.77) vs 0.58 (95% CI, 0.44-0.70) for the adenocarcinoma class. However, this outperformance was not statistically significant.

CONCLUSIONS AND RELEVANCE Results of this study suggest that the proposed attention-based deep neural network framework for BE and esophageal adenocarcinoma detection is important because it is based solely on tissue-level annotations, unlike existing methods that are based on

(continued)

Key Points

Question Can deep learning approaches accurately identify cancerous and precancerous esophagus tissue on microscopy images without training on region-of-interest annotations?

Findings In this diagnostic study of 123 histological images, a novel deep learning method, trainable without region-of-interest annotations, analyzed Barrett esophagus and esophageal adenocarcinoma whole-slide images and achieved a mean accuracy of 0.83 in classifying the test images. These results were comparable with or better than the performance from the current state-of-the-art sliding window approach, which was trained with regions of interest.

Meaning Findings of this study suggest that the proposed model eliminates the annotation bottleneck in developing deep learning models for whole-slide image analysis and is a generalizable approach to characterizing histological patterns on high-resolution microscopy images.

+ Supplemental content

Author affiliations and article information are listed at the end of this article.

Open Access. This is an open access article distributed under the terms of the CC-BY License.

Downloaded from jamanetwork.com by guest on 02/11/2025

Abstract (continued)

regions of interest. This new model is expected to open avenues for applying deep learning to digital pathology.

JAMA Network Open. 2019;2(11):e1914645. doi:10.1001/jamanetworkopen.2019.14645

Introduction

Barrett esophagus (BE) is a transformation of the normal squamous epithelium of the esophagus into metaplastic columnar epithelium.[1] Barrett esophagus is important because it predisposes patients to the increased risk of adenocarcinoma of the esophagus and gastroesophageal junction.[2,3] Compared with the general population, patients with BE have a 30 to 125 times higher risk of cancer.[4] The mean 5-year survival rate for esophageal adenocarcinoma (EAC) is less than 15% in the United States.[5] Furthermore, the incidence of EAC in the United States increased dramatically over 3 decades.[6-10] Histological diagnosis of BE requires the identification of metaplastic columnar epithelium with goblet cells (ie, intestinal metaplasia).[11] Evaluating the development of the premalignant and malignant neoplasm in BE shows a moderate interobserver variability, with a mean κ coefficient of less than 0.50 even among subspecialized gastrointestinal pathologists.[12]

In digital pathology, tissue slides are scanned as high-resolution images. High resolution is necessary because each slide contains thousands of cells, for which the cellular structures must be visible to allow the identification of regions of the tissue with diseases or lesions. The size of lesions is often relatively small, and most of the tissue areas in a given slide are normal. Even for highly trained pathologists, localizing the decisive regions of interest (ROIs) containing lesions for the classification of the whole slide is time-consuming and prone to miss an ROI.

In recent years, deep learning has made considerable advances in classifying microscopy images. The most common approach in this domain involves a sliding window model for cropped-image classification, followed by statistical methods of aggregation for whole-slide inference.[13-23] In the sliding window approach, pathologists annotate bounding boxes (ie, ROIs) on whole slides to train a classifier on small cropped images, typically in sizes ranging from 200 × 200 pixels to 500 × 500 pixels. For evaluating a whole slide, this cropped-image 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 cropped-image classification scores translates into a whole-slide diagnosis.

The sliding window approach has several limitations, however. First, given that cropped-image classifiers are needed, all images in the training set must be annotated by pathologists with bounding boxes around each ROI. Second, developing a heuristic for aggregating cropped-image classifications, which requires pathologist insight, is dependent on the nature of the classification task and is not widely scalable. Third, cropped images are classified independently of their neighbors, and whole-slide classification does not consider correlations between neighboring windows. To overcome these limitations in this study, we developed an attention mechanism that mines the ROI from high-resolution slides without explicit supervision.

Our work was inspired by attention models applied to regular-image analysis tasks, especially image captioning.[24,25] Attention mechanisms are described as a part of the prediction module that sequentially selects subsets of input to be processed.[24] Although this definition is not applicable to nonsequential tasks, the essence of attention mechanisms can be restructured for neural networks to generate a dynamic representation of features by weighting them to capture a holistic context of input. Unlike hard attention, in which an ROI is selected by a stochastic sampling process, soft attention generates a nondiscrete attention map that pays fractional attention to each region and produces better gradient flow and thus is easier to optimize. Recent advancement of soft attention enabled end-to-end training on convolutional neural network models.[26-29] For example, spatial transformer networks capture high-level information from inputs to derive affine transformation

Downloaded from jamanetwork.com by guest on 02/11/2025

parameters, which are subsequently applied to spatial invariant input for a convolutional neural network.[29] For semantic segmentation tasks, the attention mechanism is applied to learn multiscale features.[26] Residual attention networks use soft attention masks to extract features in different granularities.[28]

For analyzing images in detail, a top-down, recurrent attention, convolutional neural network has been proposed.[27] To put our work into perspective, that proposed model is based on the soft attention mechanism in feature space but is designed for the classification of high-resolution images that are not typically encountered in the field of computer vision. The attention mechanism has several applications in the medical domain, such as using soft attention to generate masks around lesion areas on computed tomography images[30] and using recurrent attention models fused with reinforcement learning to locate lung nodules[31] or enlarged hearts[32] in chest radiography images. In pathology, recorded navigation of pathologists has been used as attention maps to detect carcinoma.[33] Soft attention has been deployed in 2 parallel networks for the classification of thorax disease.[30] Although we drew inspiration from this earlier work, our proposed attention-based model is different in that it provides a novel framework to directly reuse extracted features in a single attention network.

In this study, we developed a model that uses a convolutional attention-based mechanism to classify microscopy images. This attention-based model has 3 major advantages over the existing sliding window method. First, our model dynamically identifies ROIs in a high-resolution image and makes a whole-slide classification based on the analysis of only selected regions. This process is analogous to how pathologists examine slides under the microscope. Second, our proposed model is trainable end to end with only tissue-level labels. All components of the model are optimized through backpropagation. Unlike the sliding window approach, the model does not need bounding box annotations for ROIs or pathologist insight for heuristic development. Third, our model has a flexible architecture with regard to input size for images. Inspired by fully convolutional network philosophy,[34] the model's grid-based attention module uses a 3-dimensional (3-D) convolution operation that does not require a fixed-size input grid. The input size can be any rectangular shape that fits in the memory of graphic processing units, which all modern deep learning frameworks use to accelerate computations.

Methods

Data Set

For this diagnostic study, whole-slide images were collected from patients who underwent endoscopic esophagus and gastroesophageal junction mucosal biopsy between January 1, 2016, and December 31, 2018, at Dartmouth-Hitchcock Medical Center, a tertiary academic medical center in Lebanon, New Hampshire. The use of data collected for this study was approved by the Dartmouth Institutional Review Board, which waived the requirement of informed consent as the collected data were deidentified. The study is in compliance with the Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects.[35] In addition, the study followed the Standards for Reporting of Diagnostic Accuracy (STARD) reporting guidelines.[36]

A scanner (Aperio AT2; Leica Biosystems Inc) was used to digitize hematoxylin-eosin–stained whole-slide images at 20× magnification. Scanning with 20× magnification is routinely performed in the clinical workflow for faster scanning throughput and efficient file size. We had a total of 180 whole-slide images, of which 116 (64.4%) were used as the training set and 64 (35.6%) were used as the testing set. Of the training set, 23 whole-slide images (19.8%) were reserved for validation. These whole-slide images can cover multiple pieces of tissue. Therefore, the whole-slide images were separated into 379 high-resolution images later in the preprocessing step, with each image covering a single piece of tissue.

To determine labels for whole-slide images and to train the existing state-of-the-art sliding window approach as the baseline, 2 of our expert pathologists from the Department of Pathology

Downloaded from jamanetwork.com by guest on 02/11/2025

and Laboratory Medicine at Dartmouth-Hitchcock Medical Center (A.S., B.R.) annotated bounding boxes around lesions in these images (eMethods 1 in the Supplement). We considered these labels as the reference standard, as any disagreements in annotation were resolved through further discussion among our senior domain-expert pathologist annotators. These bounding boxes were not needed in training the proposed attention-based model.

This study used categories of esophageal dysplasia and carcinoma based on the Vienna classification system.[37] The normal class included normal squamous epithelium, normal squamous and columnar junctional epithelium, and normal columnar epithelium. Barrett esophagus negative for dysplasia was included in the BE-no-dysplasia class. Barrett esophagus is defined by columnar epithelium with goblet cells (intestinal metaplasia) and preservation of orderly glandular architecture of the columnar epithelium with surface maturation. The BE-with-dysplasia class included low-grade dysplasia (noninvasive low-grade neoplasia) and high-grade dysplasia (noninvasive high-grade neoplasia). 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 demonstrated marked cytologic atypia, including loss of polarity, severe nuclear enlargement and hyperchromasia, numerous mitotic figures, and architectural abnormalities such as lateral budding, branching, and villous formation as well as variation in the size and shape of crypts.

In contrast to the Vienna classification system, we merged BE with low-grade dysplasia and high-grade classes into 1 class owing to the low number of collected samples for each class. The adenocarcinoma class included invasive carcinoma (intramucosal carcinoma and submucosal carcinoma and beyond) and high-grade dysplasia suggestive of invasive carcinoma. Cases in the adenocarcinoma class may present the following features: single-cell infiltration, sharply angulated glands, small glands in a back-to-back pattern, confluent glands, cribriform or solid growth, ulceration occurring within high-grade dysplasia, dilated dysplastic glands with necrotic debris, or dysplastic glands undermining squamous epithelium.

Two-Step Method and Testing

The proposed attention-based model has 2 steps, which are shown in **Figure 1**. The first step is the extraction of grid-based features from the high-resolution image, at which point each grid cell in the whole slide is analyzed to generate a feature map (Figure 1A and B). The second step is the application of the attention mechanism on the extracted features for slide classification (Figure 1C). The feature extractor is jointly optimized across all the grid cells with the attention module in an end-to-end fashion. In the end-to-end training pipeline, the cross-entropy loss over all classes is computed on class predictions. The loss is backpropagated to optimize all parameters in the network without any manual adjustment for attention modules. The model does not need bounding box annotations around ROIs, and all optimization is done to only the labels at the tissue level. Further details of the model architecture of the grid-based feature extraction and attention-based classification are provided in eMethods 2 in the Supplement.

To evaluate the attention-based classification model for high-resolution microscopy images, we applied the steps to high-resolution scanned slides of tissues endoscopically removed from patients who were at risk for esophageal cancer. We compared the performance results of the proposed model with those of the state-of-the-art sliding window approach.[22]

For preprocessing, we removed the white background from the slides and extracted only regions of the images that contained tissue. eFigure 1A in the Supplement shows a typical whole-slide image from the data set. These whole-slide images can cover multiple pieces of tissue, so we separated them into subimages with each covering only a single piece of tissue. The median (interquartile range) width of the tissues was 4500 (3000-6500) pixels and the median (interquartile range) height was 5500 (4000-7500) pixels. Every tissue 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

Downloaded from jamanetwork.com by guest on 02/11/2025

implication clinically. If no abnormal lesions were found in an image, it was assigned to the normal class. After this preprocessing step, each image was assigned to 1 of 4 classes: normal, BE-no-dysplasia, BE-with-dysplasia, and adenocarcinoma (eFigure 1B in the Supplement).

The data set included 379 images after preprocessing. One-third of the data set was reserved for testing. To avoid possible data leakage, we placed all tissues extracted from 1 whole-slide image into the same set of images when the training and testing sets were split. The **Table** summarizes the results of the testing set.

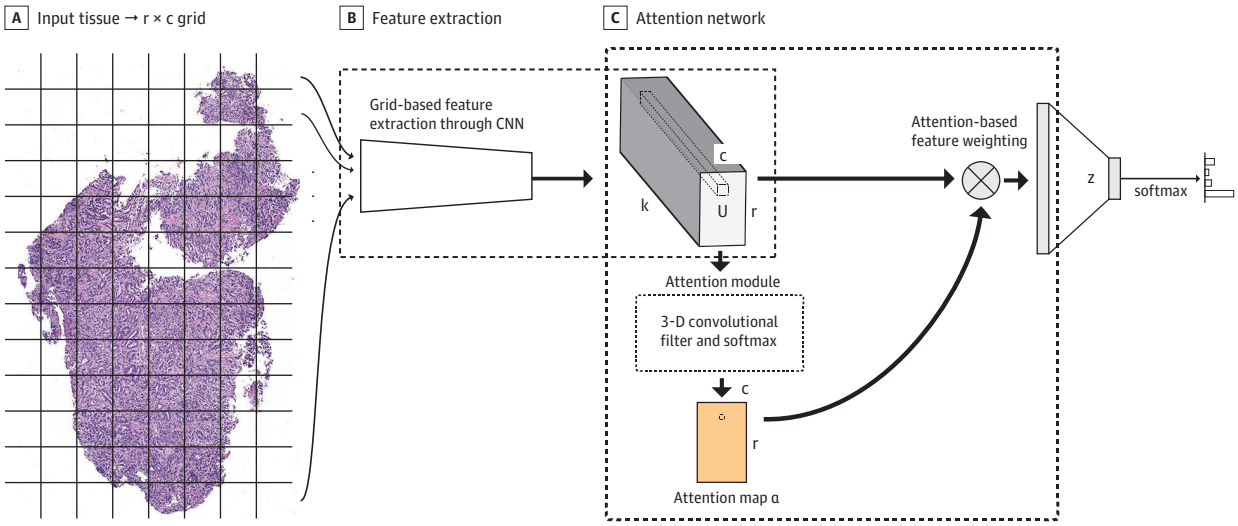
Sliding Window Approach as Baseline

To compare the proposed model with previous methods for high-resolution image analysis, we implemented the current state-of-the-art sliding window approach[22] as a baseline. For this method, we used the annotated bounding box labels to generate small, cropped images of 224 × 224 pixel size for training a cropped-image 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 MSRA (Microsoft Research Asia) initialization.[38] We optimized the model with a cross-entropy loss function for 100 epochs, using standard weight regularization techniques and learning rate decay. We trained the cropped-image classifier to predict the class of any given window on a high-resolution image. For whole-slide inference, we performed a grid search of the validation set for optimal thresholds to filter noise. Then, our 2 pathologists (A.S., B.R.) were consulted to develop heuristics for aggregating cropped-image predictions. We chose the thresholds and heuristics that performed the best on the validation set and applied those to the whole-slide images in the testing set.

Attention-Based Model

We implemented the attention-based model as described. Given the size of features extracted from the ResNet-18 model, we used 512 × 3 × 3, 3-D convolutional filters in the attention module, with implicit zero padding of 0 for depth, 1 for height, and 1 for width dimensions. We used 64 of these filters to increase the robustness of the attention module, as patterns in the feature space are likely

Figure 1. Overview of Proposed Attention-Based Model



A, An input image is divided into r × c grid cells (dividing lines are shown only for visualization). B, Features extracted from each grid cell build a grid-based feature map tensor U. C, Learnable 3-dimensional convolutional filters of size k × d × d (where d denotes the height and width of the convolutional filters) are applied on U feature map to generate an attention map α, which operates as the weights for an affine combination of feature vectors in U. The α represents a 2-dimensional attention map whose size is r in height and c in width; CNN, convolutional neural network; r and c, the number of rows and columns of input tissue grid; U, a tensor of features extracted from each grid cell, and its size is r in height, c in width, and k in depth; and z, a vector of features representing a whole-input image.

Downloaded from jamanetwork.com by guest on 02/11/2025

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[39] with $P = .50$ after concatenating all of the feature vectors. We initialized the entire network with MSRA initialization for convolutional filters,[38] unit weight and zero bias for batch normalizations,[40] and Glorot initialization for fully connected layers.[41] Only the cross-entropy loss against class labels was used in training. Other information, such as the location of bounding boxes, was not given to the network as guidance to optimal attention maps. The model identified such ROIs automatically.

We initialized the feature extraction network with weights pretrained on the ImageNet data set.[42] Input for the network was extracted grid cells of 492 × 492 pixels that were resized to 224 × 224 pixels. We normalized the input values by the mean (SD) of pixel values computed over all tissues in the training set. In training, 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, high-resolution images. For data augmentation, we applied random rotation and random scaling, with a scaling factor between 0.8 and 1.2 during

Table. Classification Results for the Testing Set[a]

Metric	Sliding Window Approach Performance (95% CI)[b]	Attention-Based Model Performance (95% CI)
Normal class		
Accuracy	0.63 (0.56-0.69)	0.70 (0.64-0.76)
Recall	0.62 (0.53-0.71)	0.69 (0.61-0.77)
Precision	0.60 (0.51-0.69)	0.68 (0.59-0.76)
Specificity	0.63 (0.57-0.72)	0.71 (0.62-0.79)
F1 score	0.61 (0.53-0.68)	0.68 (0.61-0.75)
BE-no-dysplasia class		
Accuracy	0.85 (0.80-0.89)	0.85 (0.81-0.90)
Recall	0.43 (0.31-0.56)	0.77 (0.66-0.87)
Precision	0.87 (0.73-0.97)	0.68 (0.57-0.78)
Specificity	0.98 (0.95-1.00)	0.88 (0.83-0.93)
F1 score	0.58 (0.45-0.69)	0.72 (0.63-0.80)
BE-with-dysplasia class		
Accuracy	0.72 (0.66-0.77)	0.89 (0.84-0.92)
Recall	0.36 (0.18-0.54)	0.21 (0.07-0.38)
Precision	0.16 (0.08-0.26)	0.50 (0.20-0.80)
Specificity	0.76 (0.70-0.82)	0.97 (0.94-0.99)
F1 score	0.22 (0.11-0.33)	0.30 (0.11-0.48)
Adenocarcinoma class		
Accuracy	0.87 (0.83-0.91)	0.88 (0.84-0.92)
Recall	0.52 (0.37-0.68)	0.71 (0.57-0.85)
Precision	0.65 (0.48-0.80)	0.63 (0.49-0.76)
Specificity	0.94 (0.90-0.98)	0.91 (0.87-0.95)
F1 score	0.58 (0.44-0.70)	0.67 (0.54-0.77)
Mean		
Accuracy	0.76 (0.73-0.80)	0.83 (0.80-0.86)
Recall	0.48 (0.41-0.56)	0.60 (0.53-0.66)
Precision	0.57 (0.51-0.63)	0.62 (0.53-0.71)
F1 score	0.50 (0.43-0.56)	0.59 (0.52-0.66)

Abbreviation: BE, Barrett esophagus.

[a] The proposed attention-based model's performance was assessed on the basis of accuracy, recall, precision, specificity, and F1 score. Results were rounded to 2 decimal places. The model outperformed the sliding window baseline in both accuracy and F1 score for all classes.

[b] The sliding window approach is explained in Wei et al.[22]

Downloaded from jamanetwork.com by guest on 02/11/2025

training. We used the Adam optimizer with an initial learning rate of 1×10^{-3}, decaying by 0.95 after each epoch, and reset the learning rate to 1×10^{-4} every 50 epochs in a total of 200 epochs, similar to the cyclical learning rate.[43,44] We set the mini-batch size to 2 to maximize the use of memory on the graphic processing unit (Nvidia Titan Xp; NVIDIA Corporation). The model was implemented in PyTorch.[45] At testing, the network took a mean 0.34 seconds to analyze a high-resolution image.

Statistical Analysis

Data were analyzed in October 2018. For quantitative evaluation, 4 standard metrics were used for classification under a 1-vs-rest strategy: accuracy, recall, precision, and F1 score. To estimate 95% CIs, bootstrapping was used for all metrics. The 2-tailed McNemar-Bowker test was used, and $\alpha = .05$ was considered statistically significant. Statistical analysis was carried out with SciPy, version 1.0.0 (SciPy developers).

Results

The data set contained a total of 379 histological images, of which 195 (51.5%) were in the normal class, 80 (21.1%) were in the BE-no-dysplasia class, 46 (12.1%) were in the BE-with-dysplasia class, and 58 (15.3%) were in the adenocarcinoma class. Of the independent testing set of 123 images, 58 (47.2%) normal, 30 (24.4%) BE-no-dysplasia, 14 (11.4%) BE-with-dysplasia, and 21 (17.1%) adenocarcinoma images were used to evaluate trained models and to analyze the classification performance from both quantitative and qualitative aspects. The eTable in the Supplement provides a detailed description of the data set.

The classification results on the testing set are summarized in the Table. Compared with the sliding window baseline, the proposed model achieved better accuracy and F1 score in all classes. Especially for F1 score, which is the harmonic mean of precision and recall, the attention-based model outperformed the sliding window approach by at least 8% for each class: 0.68 (95% CI, 0.61-0.75) vs 0.61 (95% CI, 0.53-0.68) for the normal class, 0.72 (95% CI, 0.63-0.80) vs 0.58 (95% CI, 0.45-0.69) for the BE-no-dysplasia class, 0.30 (95% CI, 0.11-0.48) vs 0.22 (95% CI, 0.11-0.33) for the BE-with-dysplasia class, and 0.67 (95% CI, 0.54-0.77) vs 0.58 (95% CI, 0.44-0.70) for the adenocarcinoma class. However, this outperformance was not statistically significant at the $\alpha = .05$ level with the McNemar-Bowker test.

Results of the attention-based model had the following sensitivities for recall: 0.69 (95% CI, 0.61-0.77) for the normal class, 0.77 (95% CI, 0.66-0.87) for the BE-no-dysplasia class, 0.21 (95% CI, 0.07-0.38) for the BE-with-dysplasia class, and 0.71 (95% CI, 0.57-0.85) for the adenocarcinoma class. The specificities under this model were as follows: 0.71 (95% CI, 0.62-0.79) for the normal class, 0.88 (95% CI, 0.83-0.93) for the BE-no-dysplasia class, 0.97 (95% CI, 0.94-0.99) for the BE-with-dysplasia class, and 0.91 (95% CI, 0.87-0.95) for the adenocarcinoma class. Classification accuracies of the proposed model were 0.85 (95% CI, 0.81-0.90) for the BE-no-dysplasia class, 0.89 (95% CI, 0.84-0.92) for the BE-with-dysplasia class, and 0.88 (95% CI, 0.84-0.92) for the adenocarcinoma class. The proposed model achieved a mean accuracy of 0.83 (95% CI, 0.80-0.86).

Our quantitative analysis showed the proposed model's good performance as follows: 0.68 (95% CI, 0.61-0.75) for the normal class, 0.72 (95% CI, 0.63-0.80) for the BE-no-dysplasia class, and 0.67 (95% CI, 0.54-0.77) for the adenocarcinoma class. Both the attention-based and the sliding window models, however, did not perform as well in identifying the images of BE-with-dysplasia class: 0.30 (95% CI, 0.11-0.48) for the attention-based model and 0.22 (95% CI, 0.11-0.33) for the sliding window approach. As shown in the confusion matrix in **Figure 2**, most samples of BE-with-dysplasia images that were misclassified by the attention-based model were predicted as normal tissue. This prediction was likely associated with the BE-with-dysplasia class being the least frequent category in the data set, representing only 11% of the images. For further comparison, see the receiver operating characteristic curves of both models for each class plotted in **Figure 3**. The

Downloaded from jamanetwork.com by guest on 02/11/2025

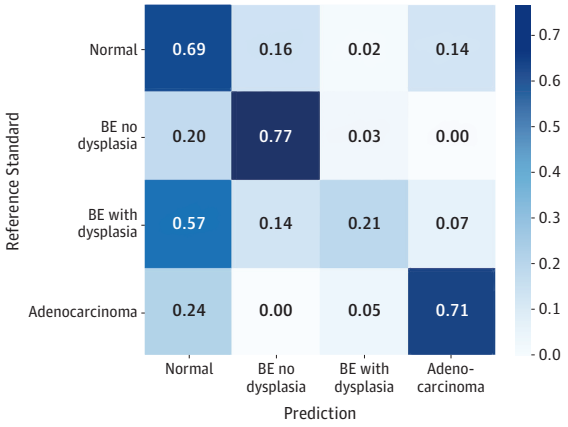
attention-based model was trained without ROI annotations yet achieved compelling area under the receiver operating characteristic curve values for each class.

The attention maps generated for all of the testing images were visualized to verify the attention mechanism in the proposed model. Characteristic examples for the adenocarcinoma class are presented in **Figure 4**. The distributions of the attention weights across different classes indicate that the attention module looks for specific features in the adenocarcinoma class (Figure 4D). For images without the target features, the attention weights are low for all regions (Figure 4A and B). In Figure 4C, the attention map is shown to be clinically on target and focused on specific regions in which BE with dysplasia progresses to adenocarcinoma as neoplastic epithelia begin to invade the muscularis mucosae.[46] eFigure 2 in the Supplement provides more examples.

Discussion

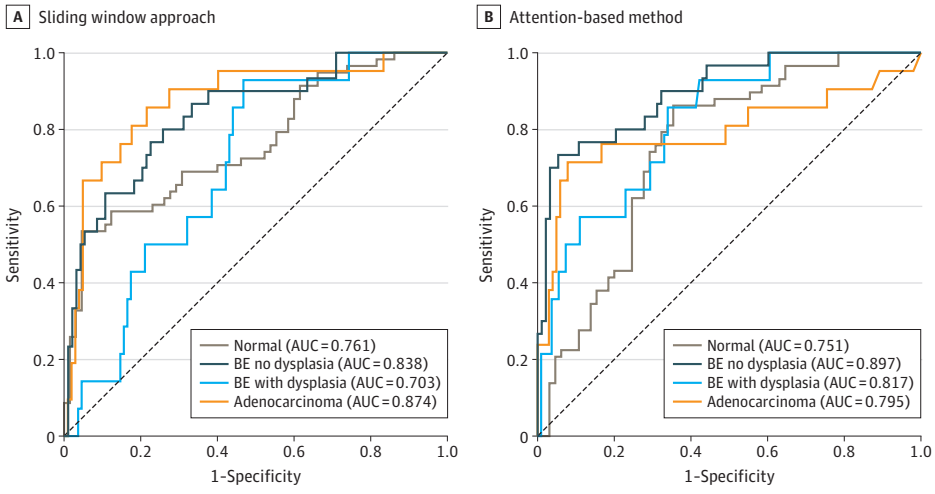
Results of this diagnostic study demonstrated the ability of attention-based deep learning architecture to detect BE or EAC. The attention-based model's classification performance on the data set was higher than that of the state-of-the-art sliding window approach. This finding is important because the proposed model needs only reference labels per tissue, whereas the existing sliding

Figure 2. Confusion Matrix for Pathologist Diagnoses and Model Predictions



The confusion matrix for different histological classes related to esophageal cancer compares the classification agreement of the attention-based model with pathologist consensus. BE indicates Barrett esophagus.

Figure 3. Performance Curves for the Sliding Window Approach and the Attention-Based Model



Receiver operating characteristic curves for the sliding window approach (A) and the proposed attention-based method (B) show the true-positive rate (y-axis) and the false-positive rate (x-axis) at various decision threshold levels for each diagnostic class. AUC indicates area under the receiver operating characteristic curve; BE, Barrett esophagus.

Downloaded from jamanetwork.com by guest on 02/11/2025

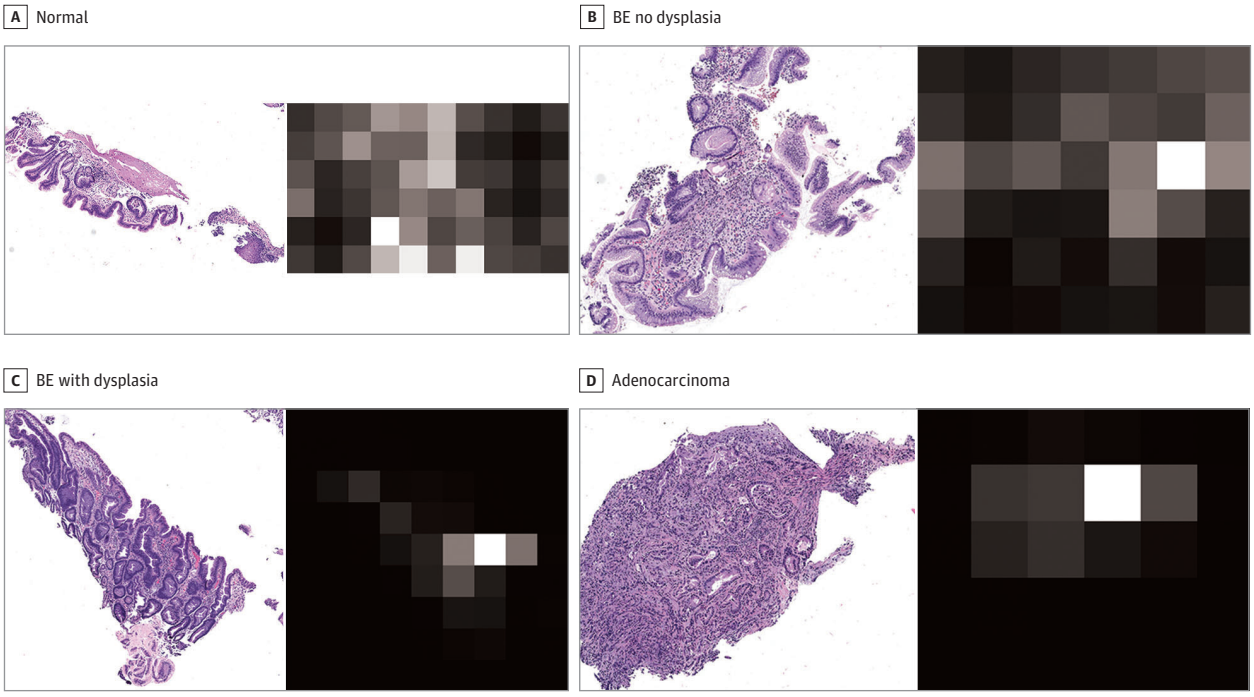
window method requires bounding box annotations for each ROI in a tissue. The higher precision in the BE-no-dysplasia and adenocarcinoma classes (ie, lower false-positives in identifying abnormality) achieved by the baseline approach may be associated with heuristic rules developed in consultation with pathologists. The rules, however, were not perfect and thus showed low recall (ie, higher false-negatives). Although both methods used a ResNet-18 model for feature extraction, the attention mechanism of the proposed model further directed the information flow and forced the network to identify local features useful for classification.

The proposed model is directly applicable to high-resolution images without resizing owing to its flexible input design. Because of the time and resources required for annotating microscopy images, having fewer requirements for these annotations would facilitate image analysis research and development. Specifically, tissue-level annotations for training the proposed architecture can potentially be retrieved through searching the pathological reports associated with microscopy images. The proposed model is potentially applicable to histological images of diseases for which training data are scarce or bounding box annotations are not available. To our knowledge, the model is the first to automate the detection of BE and EAC on histopathological slides using a deep learning approach.

Limitations

This study has some limitations. First, all experiments were conducted on slides collected from a single medical center and scanned with the same equipment. Second, the data set was relatively small compared with conventional data sets in deep learning; in particular, the number of slides of BE with dysplasia was small even after consolidating the classes of BE with low-grade dysplasia and high-grade dysplasia, resulting in lower performance for that class. To evaluate the robustness and

Figure 4. Visualization of Attention Maps Attending Adenocarcinoma Class Features



Examples of attention maps are generated by an attention module that is optimized for the features of the adenocarcinoma class. The left column shows whole-slide images from the testing set, whereas the right column shows attention maps of the selected attention module for input images from 4 classes: normal (A), Barrett esophagus (BE) no dysplasia (B), BE with dysplasia (C), and adenocarcinoma (D). Higher attention weight is denoted by white, and lower weight is denoted by black. For visualization purposes, each map is normalized so that its maximum value is 1. The accuracy of attended regions for the adenocarcinoma class images is verified qualitatively by 2 expert pathologists. In contrast, the attention module is inattentive to lower-risk-class images.

Downloaded from jamanetwork.com by guest on 02/11/2025

generalizability of the proposed model, further verification with different classification tasks and larger data sets from various institutions is required and should be pursued in future research.

Third, even with the proposed method, which was built to analyze entire tissue regions, current graphic processing units do not have enough memory capacity to process very large images. For such slides, we can divide the tissue area into manageable subtissue images. Alternatively, the feature extractor, which is the largest source of memory consumption in the proposed approach, can be optimized to address this issue. The ResNet-18 architecture used in the attention-based model achieved high performance with a relatively low number of parameters. There is, however, room for further reduction of parameters while maintaining high performance, which we intend to pursue in future studies.

Conclusions

In this diagnostic study, we developed an attention-based model for high-resolution microscopy image analysis. Analogous to how pathologists examine slides under the microscope, the model uses weighted features from the entire slide to classify microscopy images. Results showed that the model marginally outperformed the current sliding window approach on a data set of esophagus tissue with 4 classes (normal, BE-no-dysplasia, BE-with-dysplasia, and adenocarcinoma). Previous methods for analyzing microscopy images were limited by bounding box annotations and unscalable heuristics. The model presented here was trained end to end with labels only at the tissue level, thus removing the need for high-cost data annotation and creating new opportunities for applying deep learning in digital pathology.

ARTICLE INFORMATION

Accepted for Publication: September 15, 2019.

Published: November 6, 2019. doi:10.1001/jamanetworkopen.2019.14645

Open Access: This is an open access article distributed under the terms of the CC-BY License. © 2019 Tomita N et al. *JAMA Network Open*.

Corresponding Author: Saeed Hassanpour, PhD, Geisel School of Medicine, Department of Biomedical Data Science, HB 7261, Dartmouth College, Hanover, NH 03755 (saeed.hassanpour@dartmouth.edu).

Author Affiliations: Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Hanover, New Hampshire (Tomita, Abdollahi, Wei, Hassanpour); Department of Computer Science, Dartmouth College, Hanover, New Hampshire (Wei, Hassanpour); Department of Pathology and Laboratory Medicine, Dartmouth-Hitchcock Medical Center, Lebanon, New Hampshire (Ren, Suriawinata); Department of Epidemiology, Geisel School of Medicine, Dartmouth College, Hanover, New Hampshire (Hassanpour).

Author Contributions: Mr Tomita and Dr Hassanpour had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Concept and design: Tomita, Abdollahi, Wei, Suriawinata, Hassanpour.

Acquisition, analysis, or interpretation of data: Tomita, Abdollahi, Ren, Suriawinata, Hassanpour.

Drafting of the manuscript: Tomita, Abdollahi, Wei, Suriawinata, Hassanpour.

Critical revision of the manuscript for important intellectual content: All authors.

Statistical analysis: Tomita, Abdollahi.

Obtained funding: Hassanpour.

Administrative, technical, or material support: Wei, Ren, Suriawinata, Hassanpour.

Supervision: Suriawinata, Hassanpour.

Conflict of Interest Disclosures: Mr Tomita and Dr Hassanpour reported holding a pending patent to System and Method for Attention-Based Classification of High-Resolution Microscopy Images. No other disclosures were reported.

Downloaded from jamanetwork.com by guest on 02/11/2025

Funding/Support: This research was supported in part by grants R01LM012837 and P20GM104416 from the National Institutes of Health.

Role of the Funder/Sponsor: The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Additional Contributions: Lamar Moss, BA, and Maksim Bolonkin, MS, Dartmouth College, helped with the manuscript. These individuals were not compensated for their contribution.

REFERENCES

1. Haggitt RC. Barrett's esophagus, dysplasia, and adenocarcinoma. *Hum Pathol*. 1994;25(10):982-993. doi:10.1016/0046-8177(94)90057-4

2. Wild CP, Hardie LJ. Reflux, Barrett's oesophagus and adenocarcinoma: burning questions. *Nat Rev Cancer*. 2003;3(9):676-684. doi:10.1038/nrc1166

3. Conio M, Filiberti R, Blanchi S, et al; Gruppo Operativo per lo Studio delle Precancerosi Esofagee (GOSPE). Risk factors for Barrett's esophagus: a case-control study. *Int J Cancer*. 2002;97(2):225-229. doi:10.1002/ijc.1583

4. Stein HJ, Siewert JR. Barrett's esophagus: pathogenesis, epidemiology, functional abnormalities, malignant degeneration, and surgical management. *Dysphagia*. 1993;8(3):276-288. doi:10.1007/BF01354551

5. Polednak AP. Trends in survival for both histologic types of esophageal cancer in US surveillance, epidemiology and end results areas. *Int J Cancer*. 2003;105(1):98-100. doi:10.1002/ijc.11029

6. Bollschweiler E, Wolfgarten E, Gutschow C, Hölscher AH. Demographic variations in the rising incidence of esophageal adenocarcinoma in white males. *Cancer*. 2001;92(3):549-555. doi:10.1002/1097-0142(20010801)92:3<549::AID-CNCR1354>3.0.CO;2-L

7. Blot WJ. Esophageal cancer trends and risk factors. *Semin Oncol*. 1994;21(4):403-410.

8. Daly JM, Karnell LH, Menck HR. National Cancer Data Base report on esophageal carcinoma. *Cancer*. 1996;78(8):1820-1828. doi:10.1002/(SICI)1097-0142(19961015)78:8<1820::AID-CNCR25>3.0.CO;2-Z

9. Brown LM, Devesa SS. Epidemiologic trends in esophageal and gastric cancer in the United States. *Surg Oncol Clin N Am*. 2002;11(2):235-256. doi:10.1016/S1055-3207(02)00002-9

10. Edgren G, Adami H-O, Weiderpass E, Nyrén O. A global assessment of the oesophageal adenocarcinoma epidemic. *Gut*. 2013;62(10):1406-1414. doi:10.1136/gutjnl-2012-302412

11. Paull A, Trier JS, Dalton MD, Camp RC, Loeb P, Goyal RK. The histologic spectrum of Barrett's esophagus. *N Engl J Med*. 1976;295(9):476-480. doi:10.1056/NEJM197608262950904

12. Coco DP, Goldblum JR, Hornick JL, et al. Interobserver variability in the diagnosis of crypt dysplasia in Barrett esophagus. *Am J Surg Pathol*. 2011;35(1):45-54. doi:10.1097/PAS.0b013e3181ffdd14

13. Korbar B, Olofson AM, Miraflor AP, et al. Looking under the hood: deep neural network visualization to interpret whole-slide image analysis outcomes for colorectal polyps. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Piscataway, NJ: IEEE; 2017:821-827. http://openaccess.thecvf.com/content_cvpr_2017_workshops/w8/papers/Korbar_Looking_Under_the_CVPR_2017_paper.pdf. Accessed October 3, 2019.

14. Hou L, Samaras D, Kurc TM, et al. Patch-based convolutional neural network for whole slide tissue image classification. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: IEEE; 2016:2424-2433. https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Hou_Patch-Based_Convolutional_Neural_CVPR_2016_paper.pdf. Accessed October 3, 2019.

15. Cosatto E, Laquerre P-F, Malon C, et al. Automated gastric cancer diagnosis on h&e-stained sections; training a classifier on a large scale with multiple instance machine learning. Medical Imaging 2013. In: *Proceedings Volume 8676, Medical Imaging 2013: Digital Pathology*. Lake Buena Vista, FL: SPIE Medical Imaging; 2013. http://malon.info/papers/spiemidp13-gastric-mil.pdf. Accessed October 3, 2019.

16. Saha M, Chakraborty C, Racoceanu D. Efficient deep learning model for mitosis detection using breast histopathology images. *Comput Med Imaging Graph*. 2018;64:29-40. doi:10.1016/j.compmedimag.2017.12.001

17. Komura D, Ishikawa S. Machine learning methods for histopathological image analysis. *Comput Struct Biotechnol J*. 2018;16:34-42. doi:10.1016/j.csbj.2018.01.001

18. Xu J, Luo X, Wang G, Gilmore H, Madabhushi A. A Deep Convolutional Neural Network for segmenting and classifying epithelial and stromal regions in histopathological images. *Neurocomputing*. 2016;191:214-223. doi:10.1016/j.neucom.2016.01.034

Downloaded from jamanetwork.com by guest on 02/11/2025

19. Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat Med*. 2018;24(10):1559-1567. doi:10.1038/s41591-018-0177-5

20. Liu Y, Gadepalli K, Norouzi M, et al. Detecting cancer metastases on gigapixel pathology images. arXiv. Preprint posted March 3, 2017. Revised March 8, 2017.

21. Wei JW, Wei JW, Jackson CR, Ren B, Suriawinata AA, Hassanpour S. Automated detection of celiac disease on duodenal biopsy slides: a deep learning approach. *J Pathol Inform*. 2019;10(1):7-7. doi:10.4103/jpi.jpi_87_18

22. Wei JW, Tafe LJ, Linnik YA, Vaickus LJ, Tomita N, Hassanpour S. Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. *Sci Rep*. 2019;9(1):3358. doi:10.1038/s41598-019-40041-7

23. Korbar B, Olofson AM, Miraflor AP, et al. Deep learning for classification of colorectal polyps on whole-slide images. *J Pathol Inform*. 2017;8:30. doi:10.4103/jpi.jpi_34_17

24. Xu K, Ba J, Kiros R, et al. Show, attend and tell: neural image caption generation with visual attention. In: Proceedings of the 32nd International Conference on Machine Learning; Lille, France. *JMLR*. 2015;37:2048-2057. http://proceedings.mlr.press/v37/xuc15.pdf. Accessed October 3, 2019.

25. You Q, Jin H, Wang Z, Fang C, Luo J. Image captioning with semantic attention. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: IEEE; 2016:4651-4659. http://openaccess.thecvf.com/content_cvpr_2016/papers/You_Image_Captioning_With_CVPR_2016_paper.pdf. Accessed October 3, 2019.

26. Chen L-C, Yang Y, Wang J, Xu W, Yuille AL. Attention to scale: scale-aware semantic image segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: IEEE; 2016:3640-3649. http://openaccess.thecvf.com/content_cvpr_2016/papers/Chen_Attention_to_Scale_CVPR_2016_paper.pdf. Accessed October 3, 2019.

27. Fu J, Zheng H, Mei T. Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Piscataway, NJ: IEEE; 2017;2:4476-4484. http://openaccess.thecvf.com/content_cvpr_2017/papers/Fu_Look_Closer_to_CVPR_2017_paper.pdf. Accessed October 3, 2019.

28. Wang F, Jiang M, Qian C, et al. Residual attention network for image classification. arXiv. Preprint posted April 23, 2017.

29. Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks. *Adv Neural Inf Process Syst*. 2015;28:2017-2025. https://pdfs.semanticscholar.org/84f1/3b5dedcbb84401a85abd1267c0f3e6c71d2a.pdf. Accessed October 3, 2019.

30. Guan Q, Huang Y, Zhong Z, Zheng Z, Zheng L, Yang Y. Diagnose like a radiologist: attention guided convolutional neural network for thorax disease classification. arXiv. Preprint posted January 30, 2018.

31. Pesce E, Ypsilantis P-P, Withey S, Bakewell R, Goh V, Montana G. Learning to detect chest radiographs containing lung nodules using visual attention networks. arXiv. Preprint posted December 4, 2017.

32. Ypsilantis P-P, Montana G. Learning what to look in chest X-rays with a recurrent visual attention model. arXiv. Preprint posted January 23, 2017.

33. Corredor G, Whitney J, Arias V, Madabhushi A, Romero E. Training a cell-level classifier for detecting basal-cell carcinoma by combining human visual attention maps with low-level handcrafted features. *J Med Imaging (Bellingham)*. 2017;4(2):021105. doi:10.1117/1.JMI.4.2.021105

34. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: IEEE; 2015:3431-3440. https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Long_Fully_Convolutional_Networks_2015_CVPR_paper.pdf. Accessed October 3, 2019.

35. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA*. 2013;310(20):2191-2194. doi:10.1001/jama.2013.281053

36. Bossuyt PM, Reitsma JB, Bruns DE, et al; STARD Group. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Radiology*. 2015;277(3):826-832. doi:10.1148/radiol.2015151516

37. Schlemper RJ, Riddell RH, Kato Y, et al. The Vienna classification of gastrointestinal epithelial neoplasia. *Gut*. 2000;47(2):251-255. doi:10.1136/gut.47.2.251

38. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: IEEE; 2015:1026-1034. https://www.cv-foundation.org/openaccess/content_iccv_2015/papers/He_Delving_Deep_into_ICCV_2015_paper.pdf. Accessed October 3, 2019.

Downloaded from jamanetwork.com by guest on 02/11/2025

39. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*. 2014;15(1):1929-1958. https://pdfs.semanticscholar.org/6c8b/30f63f265c32e26d999aa1fef5286b8308ad.pdf. Accessed October 3, 2019.

40. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv. Preprint posted February 11, 2015. Revised March 2, 2015.

41. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics; Sardinia, Italy. *JMLR*. 2010;9:249-256. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.207.2059&rep=rep1&type=pdf. Accessed October 3, 2019.

42. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Adv Neural Inf Process Syst*. 2012;25:1097-1105. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf. Accessed October 3, 2019.

43. Smith LN. Cyclical learning rates for training neural networks. In: *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Piscataway, NJ: IEEE; 2017:464-472. https://www.nrl.navy.mil/itd/aic/sites/www.nrl.navy.mil.itd.aic/files/pdfs/CLR%20%28Smith%203-23-17%29.pdf. Accessed October 3, 2019.

44. Loshchilov I, Hutter F. Sgdr: stochastic gradient descent with warm restarts. Paper presented at: 2017 Fifth International Conference on Learning Representations; April 24-26, 2017; Toulon, France.

45. Paszke A, Gross S, Chintala S, Chanan G. PyTorch. https://pytorch.org/. Accessed October 3, 2019.

46. Odze RD, Goldblum JR. *Odze and Goldblum Surgical Pathology of the GI Tract, Liver, Biliary Tract and Pancreas*. 3rd ed. Philadelphia, PA: Saunders Elsevier; 2014.

SUPPLEMENT.

eMethods 1. Details of Our Data Annotation Procedure
eMethods 2. Details of Our Attention-Based Deep Learning Architecture
eFigure 1. Typical Examples of a Whole-Slide Image and Class-Associated Patches
eFigure 2. Additional Examples of Visualized Attention Maps Attending to Adenocarcinoma Class Features
eTable. Class Distribution of Images in Our Dataset
eReferences.

Downloaded from jamanetwork.com by guest on 02/11/2025