Original Research

Automated Detection of Celiac Disease on Duodenal Biopsy Slides: A Deep Learning Approach

Jason W. Wei[1,2], Jerry W. Wei[1], Christopher R. Jackson[3], Bing Ren[3], Arief A. Suriawinata[3], Saeed Hassanpour[1,2,4]

Departments of [1]Biomedical Data Science and [2]Computer Science, Dartmouth College, Hanover, [3]Department of Pathology and Laboratory Medicine, Dartmouth-Hitchcock Medical Center, Lebanon, [4]Department of Epidemiology, Dartmouth College, Hanover, New Hampshire, USA

Received: 15 November 2018 Accepted: 31 January 2019 Published: 08 March 2019

Abstract

Context: Celiac disease (CD) prevalence and diagnosis have increased substantially in recent years. The current gold standard for CD confirmation is visual examination of duodenal mucosal biopsies. An accurate computer-aided biopsy analysis system using deep learning can help pathologists diagnose CD more efficiently. **Subjects and Methods:** In this study, we trained a deep learning model to detect CD on duodenal biopsy images. Our model uses a state-of-the-art residual convolutional neural network to evaluate patches of duodenal tissue and then aggregates those predictions for whole-slide classification. We tested the model on an independent set of 212 images and evaluated its classification results against reference standards established by pathologists. **Results:** Our model identified CD, normal tissue, and nonspecific duodenitis with accuracies of 95.3%, 91.0%, and 89.2%, respectively. The area under the receiver operating characteristic curve was >0.95 for all classes. **Conclusions:** We have developed an automated biopsy analysis system that achieves high performance in detecting CD on biopsy slides. Our system can highlight areas of interest and provide preliminary classification of duodenal biopsies before review by pathologists. This technology has great potential for improving the accuracy and efficiency of CD diagnosis.

Keywords: Celiac disease, deep learning, digital pathology, duodenal biopsy, whole-slide imaging

Introduction

Celiac disease (CD), an autoimmune disorder triggered from the consumption of gluten, affects as much as 1% of the population worldwide.[1,2] Patients who are diagnosed with CD undergo treatment in the form of a lifelong gluten-free diet, which requires substantial patient education, motivation, and follow-up.[3] Recent studies have found that the prevalence of CD has increased dramatically in the United States and Europe and that undiagnosed CD was associated with a nearly four-fold increase in risk of death.[4-6] In fact, CD remains undiagnosed in the majority of affected people, highlighting the need for more frequent and accurate methods for its detection.[7-9]

CD diagnosis involves serological testing of celiac-specific antibodies, followed by microscopic examination of duodenal biopsies, which are considered the gold standard in diagnostic confirmation of CD.[10,11] Typically, four-to-six duodenal samples are taken from the patient by an endoscopic procedure, and these samples are visually examined by a pathologist. A confirmatory diagnosis requires detection of histological changes associated with the disease, which are classified according to the guidelines from either Marsh,[12] Marsh modified (Oberhuber),[13] or Corazza *et al*.[14] Endoscopic findings that indicate CD include scalloped folds with or without mosaic pattern mucosa, reduction in the number of folds, and nodular mucosa.[15] The spectrum of histologic changes in CD ranges from only increasing intraepithelial lymphocytes with preserved villous architecture to mild villous blunting to complete villous atrophy.[16] Studies have shown that the histological diagnosis of biopsies is subject to a significant degree of interobserver variability.[14,17-19] One potential method for improving the accuracy of CD detection on duodenal biopsies is to apply automated image analysis

Address for correspondence: Dr. Saeed Hassanpour, 1 Medical Center Drive, HB 7261, Lebanon, New Hampshire 03756, USA.
E-mail: saeed.hassanpour@dartmouth.edu

Access this article online

Quick Response Code:

Website: www.jpathinformatics.org

DOI: 10.4103/jpi.jpi_87_18

How to cite this article: 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;10:7.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2019/10/1/7/253722

to aid pathologists. Since the prevalence of CD is increasing, active case finding is currently being used to screen more patients.[20,21] An automated biopsy analysis system could help pathologists by filtering and prepopulating scans, improving efficiency and turnaround time.

Recently, a subfield in artificial intelligence known as deep learning has produced a set of image analysis techniques that automatically extract relevant features, transforming the field of computer vision.[22] Deep neural networks use a data-driven approach to learn multilevel representations of data, allowing for comprehensive image analysis and classification.[23] These techniques are being increasingly applied to medical imaging to assist radiologists and pathologists.[24] In gastroenterology, the previous studies have already used deep neural networks to classify colorectal polyps on biopsy and colonoscopy images,[25-27] intraductal papillary mucinous neoplasms in magnetic resonance images,[28] and diabetic retinopathy in retinal fundus photographs.[29] For CD in particular, large video datasets captured during endoscopies have facilitated quantitative analysis with deep learning.[30,31] However, endoscopic classification is for the most part not used for confirming the diagnosis of CD. In this study, we developed a deep learning model that detects CD from duodenal biopsy images, the gold standard for diagnosis. We evaluated our model on an independent test set of 212 whole-slide images.

Subjects and Methods

Data collection

To train and evaluate our model for CD detection, we collected whole-slide images from all patients who underwent duodenal biopsies from 2016 to 2018 at the Dartmouth-Hitchcock Medical Center (DHMC), a tertiary academic care center in Lebanon, NH. These slides contain hematoxylin and eosin stained formalin-fixed paraffin-embedded tissue specimens and were scanned by a Leica Aperio whole-slide scanner at ×20 magnification by the Department of Pathology and Laboratory Medicine at DHMC. In total, we collected 1230 slides from 1048 patients. We randomly partitioned 1018 of these whole-slide images from 681 patients for model training and 212 whole-slide images from 163 patients as an independent test set for the final evaluation of our model. There was no patient overlap for the slides in the training and test sets.

Slide annotation

All whole-slide images used in our study were diagnosed by attending pathologists on gastrointestinal pathology service at the time as either normal, CD, or nonspecific duodenitis. Normal duodenal biopsies show preserved villous architecture with no mucosal injury or acute or chronic inflammation. CD biopsies show a spectrum of histologic changes as described in Marsh classification,[12] including partial to total villous atrophy with intraepithelial lymphocytosis, chronic inflammation, and crypt regenerative hyperplasia. Nonspecific duodenitis includes histologic changes including peptic duodenitis, drug-induced injury, and various other differential diagnoses of villous atrophy and acute and chronic inflammation. These labels were parsed from the medical record database and assigned as reference standard for slides used during model training. The training slides were then further split into a training set of 790 images and a development set of 228 images. Training set slides were used for training our neural network, while development set images were used for hyperparameter tuning. For the independent test set of 212 images, however, all labels were separately reviewed and confirmed by two gastrointestinal pathologists. Disagreements between original labels and new

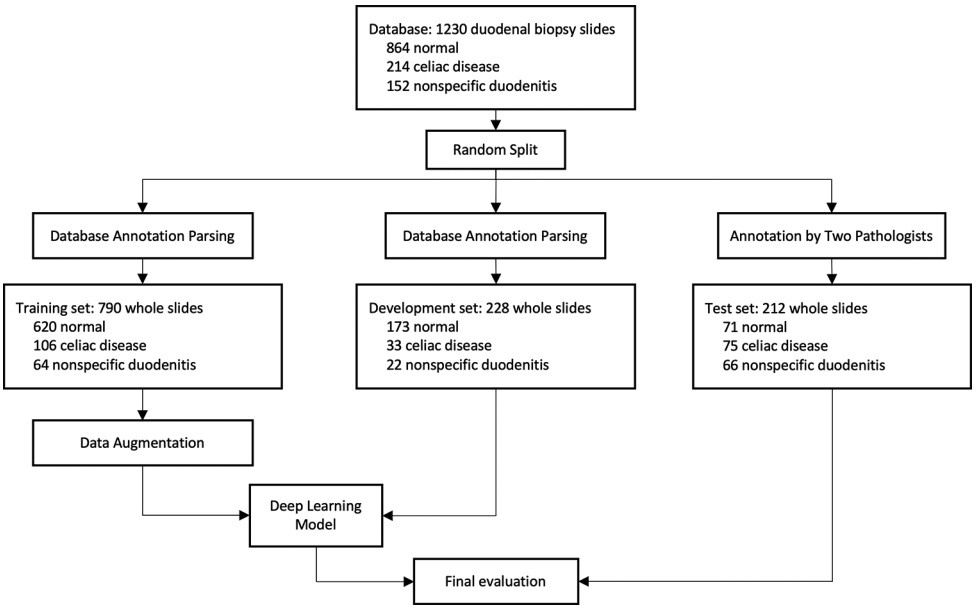

Figure 1: Data flow diagram for allocating whole slides for training, development, and testing of our model. For training, patches were generated using the sliding window algorithm to train our residual network patch classifier. The development set was used to fine-tune hyperparameters and thresholds of our neural network. Finally, we evaluated our model on the test set of 212 whole-slide images with reference labels

labels were reviewed by a senior gastrointestinal pathologist, who determined final classifications. The class distributions and roles in the data flow for our training, development, and test set are shown in Figure 1.

Model development

In recent years, research in deep learning has demonstrated successful application of convolutional neural networks for image classification, including medical image analysis. In our study, we used the deep residual network (ResNet),[32] a neural network architecture built from residual blocks. ResNet significantly outperforms early deep learning models such as AlexNet[33] and VGG[34] and achieved state-of-the-art performance on the ImageNet and COCO image recognition benchmarks.[35,36] We implemented ResNet to take in square patches as inputs and output a prediction probability for each of the three classes: normal, CD, and nonspecific duodenitis.

For model training, we used a sliding window method on each high-resolution whole-slide image to generate small patches of size 224 × 224 pixels. Since some classes had more whole-slide images than others, we generated patches with different overlapping areas for each class. When inputting a patch into the model for training, we normalized the red, green, blue color channels to the mean and standard deviation of the entire training set to neutralize differences in color among slides. Then, we performed color jittering on the brightness, contrast, saturation, and hue of each patch. Finally, we randomly rotated and flipped the images across the horizontal and vertical axes. In total, we generated 80,000 patches for each of our three classes, which were then uniquely augmented during each epoch in training.

In terms of model parameters, we initialized ResNet-50, the fastest ResNet with three-layer residual blocks, with weights from the He initialization.[37] We trained our ResNet model by optimizing on a multiclass cross-entropy loss function for forty epochs on the augmented training set, starting with an initial learning rate of 0.0001 and decaying by a factor of 0.85 every epoch. We used the Adam Optimizer[38] and weight decay regularization (L2 penalty)[39] of 0.0001. Total training time was 12 h on a Titan Xp graphics processing unit (GPU).

Whole-slide inference

In whole-slide inference, we aimed to classify each whole-slide image as either normal, CD, or nonspecific duodenitis. The model is trained to classify small patches rather than entire slides, so we again used the sliding window algorithm to break down each whole slide into a collection of patches, each overlapping by one-third area. Next, we applied our trained ResNet model to classify each patch, and we filtered out noise using thresholding to discard predictions of low confidence. Given the distribution of patch predictions, we used the following heuristic to determine the whole-slide class: if more than γ patches were classified as nonspecific duodenitis, then the whole slide was classified as nonspecific duodenitis. Otherwise, the most commonly predicted class was chosen as the whole-slide prediction. Thresholds for filtering noise, as well as γ, were optimized by performing a grid search over the development set. This allowed for accurate classification of slides with a significant amount of nonspecific duodenitis that was not covering the majority of the specimen area. Figure 2 depicts the whole-slide inference process. Inference time for a single whole-slide image was about 15 s on a single Titan Xp GPU.

Evaluation and visualization

For the final evaluation, we applied our model to the independent test set of 212 whole-slide images. We compared the predictions of our model with reference standards established by pathologists and measured accuracy, precision, recall, and F1 score for each class. We calculated confidence intervals for all performance metrics using the Clopper and Pearson method.[40] In addition, we plotted receiver operating characteristic (ROC) curves and calculated area under the curve (AUC) for each class.

Furthermore, we visualized our model's predictions at both the whole-slide and patch level. At the whole-slide level, we overlaid color-coded dots on patches for which the model predicted a particular pattern. This helps pathologists quickly identify regions of the slide containing abnormal tissue. At the patch level, we used the class activation mapping (CAM) method[41] to generate a pixel-level heat map that highlights the most informative regions of the image relevant to the predicted

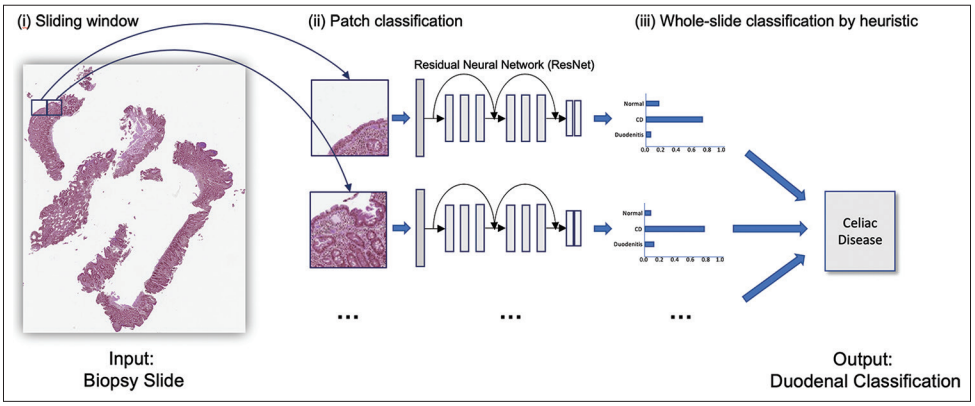

Figure 2: Overview of detection of celiac disease on whole-slide biopsy images. We used a sliding window approach on a whole-slide image to generate patches, classified each patch with a residual network model, and used a heuristic on the aggregated patch predictions to classify the whole slide

class. This demystified our classification method for each patch by revealing the most significant histologic features on the patch for each class for our model.

Results

For model selection, we validated our neural network model on the development set of 228 images. We found the optimal thresholds for filtering out noise at the patch level to be 0.7 for the normal class, 0.8 for the CD class, and 0.85 for nonspecific duodenitis. For selection of the γ threshold for percent area needed to classify a whole slide as nonspecific duodenitis, we considered our gastrointestinal pathologist's subjective examination of our model's predictions in addition to a grid search to arrive at γ=0.25. After threshold optimization, our best model applied to the development set achieved an accuracy of 95.6% for normal, 98.7% for CD, and 94.3% for nonspecific duodenitis after threshold optimization.

Performance of our model on the independent test is shown in Table 1, which includes accuracy, precision, recall, and F1 score with 95% confidence intervals. Notably, our model detects the presence of CD with an accuracy of 95.3% and an F1 score of 93.5%. Table 2 shows the confusion matrix for predicted labels versus reference labels. ROC curves and AUC for each class are shown in Figure 3. AUC was >0.95 for all classes. Figure 4 depicts whole-slide visualizations of 12 biopsy samples using dots to indicate predicted patch labels. Finally, CAM visualizations of individual patches are shown in Figure 5 to highlight relevant features used in our model's classification process. A subjective qualitative investigation of these visualizations by a gastrointestinal pathologist confirmed that the predictions of our model are generally on target.

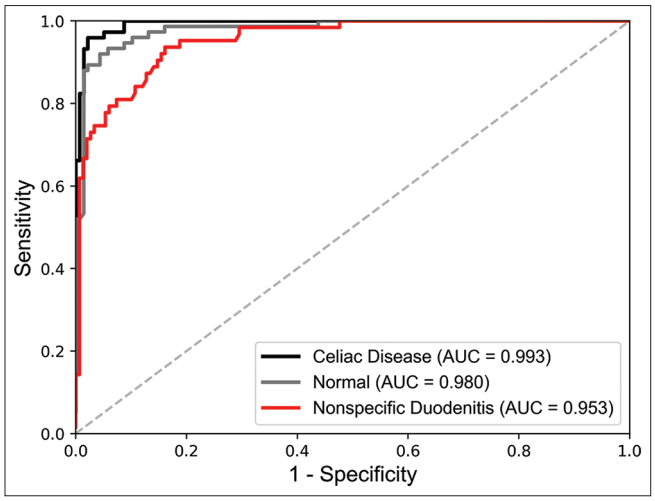

Figure 3: Receiver operating characteristic curves and their area under the curve for our model's classifications on the independent test set of 212 whole-slide biopsy images

Discussion

Duodenal biopsies are the gold standard for confirming the diagnosis of CD. The prevalence of CD has increased dramatically in recent years and active case-finding calls for more serological tests and duodenal biopsies. Detection of CD on these biopsies could potentially be enhanced and facilitated by automated image processing. In this study, we presented a deep learning model that classifies duodenal tissue and highlights the associated features and regions of interest. Previous work has used deep learning to detect CD from endoscopic images.[30,31] While acknowledging the substantive work of these investigators, endoscopies are for the most part not used for confirming CD diagnoses. Our model not only detects CD on duodenal biopsies but it also visualizes regions of normal tissue, CD, and nonspecific duodenitis to aid review by pathologists. Of note, we are not aware of any other existing system for CD detection on biopsy images.

Our model achieved high performance for detection of CD. On the independent test set of 212 images, our model detected CD with a considerable F1 score of 93.5% and AUC of 0.993. Since our model made predictions at the patch level and then aggregated them for whole-slide inference, it was relatively unaffected by noise and achieved high accuracy. For normal and nonspecific duodenitis, F1 scores were 87.2% and 81.0%, respectively. Identification of nonspecific duodenitis was more challenging because this class's tissue often also contains some normal tissue fragments, which complicate the analysis. Sixteen slides were misclassified between normal and nonspecific duodenitis, and pathologist evaluation of these errors revealed errors related to tissue orientation, fixation artifact, and patchy histologic changes. In addition, seven slides of nonspecific duodenitis were identified as CD due to focal increase of intraepithelial lymphocytes and partial villous atrophy. Performance measures for the nonspecific duodenitis class were the lowest across the board. There are several reasons for this. One could be that nonspecific duodenitis had the lowest number of training samples, comprising only 64 images in the training set compared to 620 and 106 for

Table 1: Performance of our final model for celiac disease detection on 212 duodenal biopsy whole-slide images in our test set

	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
Normal (*n*=71)	91.0 (87.2-94.9)	83.3 (75.1-91.6)	91.5 (85.1-98.0)	87.2 (79.9-95.2)
Celiac Disease (*n*=74)	95.3 (92.4-98.1)	90.0 (83.4-96.6)	97.3 (93.6-99.9)	93.5 (87.8-99.3)
Nonspecific Duodenitis (*n*=67)	89.2 (85.0-93.3)	90.7 (83.0-98.5)	73.1 (62.5-83.7)	81.0 (71.5-90.5)
Average	87.7 (83.3-92.2)	88.0 (80.5-95.4)	87.3 (79.5-95.2)	87.2 (79.4-95.1)

95% CI are shown in parentheses. CI: Confidence intervals

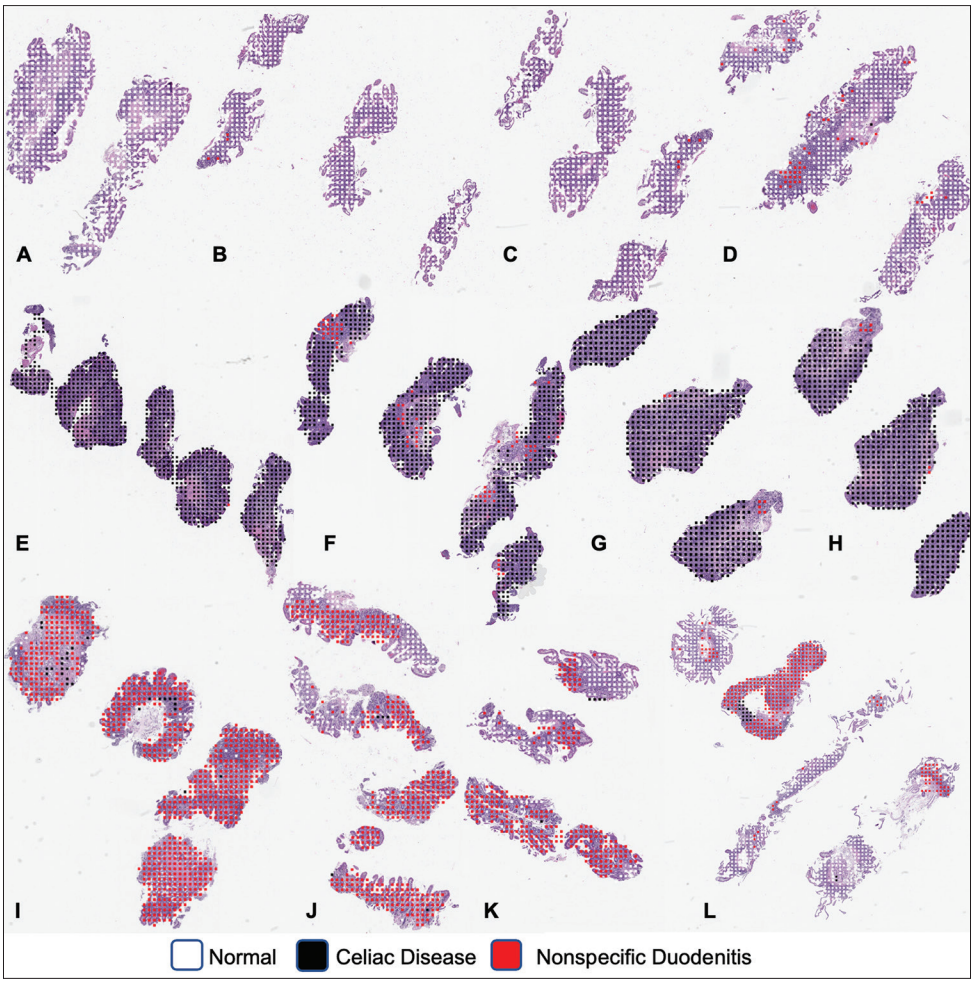

Figure 4: Visualization of patch predictions of our model at the whole-slide level (a-d) was correctly classified as normal, (e-h) was correctly classified as celiac disease, and (i-l) was correctly classified as nonspecific duodenitis

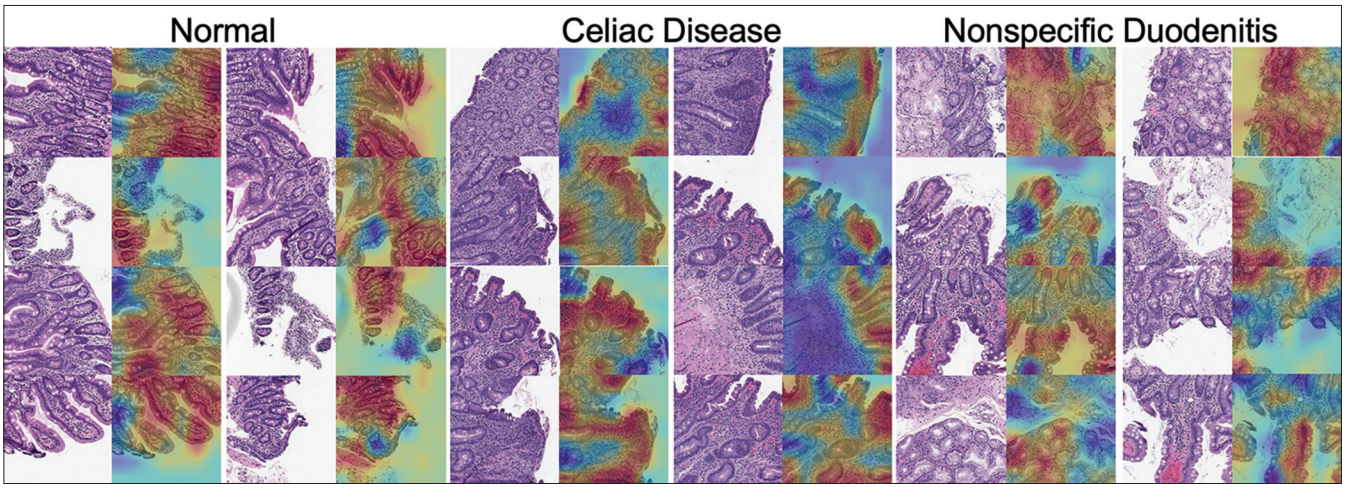

Figure 5: Class activation mapping heat maps highlighting the most informative regions of patches relevant to normal, celiac disease, and nonspecific duodenitis classes. Red regions indicate areas of attention for our residual neural network

normal tissue and CD, respectively. Another could be that this category comprises several disease entities including peptic chronic duodenitis, active duodenitis, and other nonspecific reactive changes, making it more challenging to detect since there was a wider range of histologic attributes to learn. Finally, slides labeled as nonspecific duodenitis often contained some portions of normal tissue, and since we extracted patch labels based on whole-slide labels during the training process, it is

Table 2: Confusion matrix of our final model for celiac disease detection on 212 duodenal biopsy whole-slide images in our test set

Prediction	Reference		
	Normal	CD	Nonspecific Duodenitis
Normal	65	2	11
CD	1	72	7
Nonspecific duodenitis	5	0	49

CD: Celiac disease

likely that some mislabeled data was used in model training and made it harder to detect this class in our approach.

In terms of visualization, the CAM results of our model's selected areas of attention indicate that our model has learned the correct histologic features for each class. Classification of normal tissue tended to be holistic, with attention to almost all tissue area including normal villous architecture. For CD, the model correctly identified villous atrophy, intraepithelial lymphocytosis, and chronic inflammation in the lamina propria. In the case of nonspecific duodenitis, the model identified villous thickening, Brunner's gland hyperplasia, foveolar metaplasia, and chronic inflammation.

Our results indicate that deep neural networks have substantial potential to aid gastrointestinal pathologists in diagnosing CD. For application in a clinical setting, our model could be integrated into existing laboratory information management systems to prepopulate patch predictions on slides and provide preliminary diagnoses before review by pathologists. In addition, a visualization of the slide evaluated by our model at the piecewise level could highlight precise tissue area containing abnormal or sprue patterns, allowing pathologists to quickly examine regions of interest. As CD prevalence has increased dramatically in recent years, more serological screenings and duodenal biopsies are being done for patients at risk.[20,21] With biopsies as the gold standard for diagnosis, our work aims to provide pathologists with a tool for more accurate and efficient detection of CD.

The model presented in this study is rooted in solid deep learning methodology and achieves commendable performance, but there are several limitations of our study. One limitation is that all biopsy slides were collected from a single medical center and scanned with the same equipment, so our data may not be representative of the entire range of histologic patterns in patients worldwide. Although our whole-slide scans are high resolution and we were able to extract a large number of patches for training with the sliding window method, our dataset is still small in comparison to conventional datasets in deep learning, which contains more than ten thousand unique samples per class[42,43] and more than a million unique images in total.[44] Overfitting is unlikely because we generated a large number of small patches for training and conducted final evaluation on an independent test set, but it is still a possibility. Collecting more data in collaboration with another medical center in future work would allow us to train a more generalizable neural network and could also improve our model's performance in classifying nonspecific duodenitis.

Moving forward, more work will be done to further the capabilities of our model and evaluate its use a clinical setting. Collecting an annotated dataset with specific histopathological classifications of CD and labeled bounding boxes around lesions would allow our model to classify and locate specific lesion types, providing pathologists with more comprehensive slide analysis, particularly for the nonspecific duodenitis class. Furthermore, once more data is collected, we can predict slide level results using patch predictions to train a traditional machine learning classifier such as a support vector machine or random forest, which may yield better results than our current thresholding method. In terms of clinical application, we plan on validating our model on a larger test set from multiple institutions and deploying a trial implementation of our model into laboratory information management systems at the DHMC to measure its ability to improve CD detection accuracy and efficiency. However, widespread clinical implementation of such artificial intelligence tools will require major future steps, which our group will be undertaking. Any deep learning models for computer-aided pathology must be thoroughly validated through clinical trials and be proven to enhance outcomes. Such model must also not impact the established workflow of pathologists or slow down the speed of existing programs. Most importantly, deep learning models must be accurate and gain the confidence of physicians, patients, and the medical community. In its current state, artificial intelligence has the ability to analyze images and make preliminary classifications, but much more work must be done before patients and physicians will be able to trust computers to make medical decisions. We believe that the work presented in this study is a preliminary step in this direction.

Conclusions

We have demonstrated that deep learning can achieve high accuracy in detecting CD in duodenal biopsies. Our model uses a state-of-the-art residual neural network architecture for whole-slide classification and achieved exemplary results on an independent test set of 212 whole-slide images. As CD prevalence and screening increases, we expect our model could assist pathologists in more accurate and efficient evaluation of duodenal biopsy slides.

Acknowledgment

The authors would like to thank Matthew Suriawinata for assistance with slide scanning and Sophie Montgomery and Lamar Moss for their feedback on the manuscript.

Financial support and sponsorship

This research was financially supported by the Kaminsky Research Fund from Dartmouth College and a National Institutes of Health grant, P20GM104416.

Conflicts of interest

There are no conflicts of interest.

References

1. Green PH, Cellier C. Celiac disease. N Engl J Med 2007;357:1731-43.
2. Mustalahti K, Catassi C, Reunanen A, Fabiani E, Heier M, McMillan S, *et al.* The prevalence of celiac disease in Europe: Results of a centralized, international mass screening project. Ann Med 2010;42:587-95.
3. Rubio-Tapia A, Hill ID, Kelly CP, Calderwood AH, Murray JA; American College of Gastroenterology. ACG clinical guidelines: Diagnosis and management of celiac disease. Am J Gastroenterol 2013;108:656-76.
4. Rubio-Tapia A, Kyle RA, Kaplan EL, Johnson DR, Page W, Erdtmann F, *et al.* Increased prevalence and mortality in undiagnosed celiac disease. Gastroenterology 2009;137:88-93.
5. Godfrey JD, Brantner TL, Brinjikji W, Christensen KN, Brogan DL, Van Dyke CT, *et al.* Morbidity and mortality among older individuals with undiagnosed celiac disease. Gastroenterology 2010;139:763-9.
6. Dubé C, Rostom A, Sy R, Cranney A, Saloojee N, Garritty C, *et al.* The prevalence of celiac disease in average-risk and at-risk Western European populations: A systematic review. Gastroenterology 2005;128:S57-67.
7. West J, Logan RF, Hill PG, Lloyd A, Lewis S, Hubbard R, *et al.* Seroprevalence, correlates, and characteristics of undetected coeliac disease in England. Gut 2003;52:960-5.
8. Rostami K, Mulder CJ, Werre JM, van Beukelen FR, Kerchhaert J, Crusius JB, *et al.* High prevalence of celiac disease in apparently healthy blood donors suggests a high prevalence of undiagnosed celiac disease in the Dutch population. Scand J Gastroenterol 1999;34:276-9.
9. Rubio-Tapia A, Murray JA. Classification and management of refractory coeliac disease. Gut 2010;59:547-57.
10. Green PH, Rostami K, Marsh M. Diagnosis of coeliac disease. Best Pract Res Clin Gastroenterol 2005;19:389-400.
11. Bryne G, Feighery CF. Celiac disease: Diagnosis. Methods Molecular Biology. Vol. 1326. New York, NY: Humana Press; 2015.
12. Marsh MN. Gluten, major histocompatibility complex, and the small intestine: a molecular and immunobiotic approach to the spectrum of Gluten sensitivity ('celiac sprue'). Gastroenterology 1992;102:330-54.
13. Oberhuber G. Histopathology of celiac disease. Biomed Pharmacother 2000;54:368-72.
14. Corazza GR, Villanacci V, Zambelli C, Milione M, Luinetti O, Vindigni C, *et al.* Comparison of the interobserver reproducibility with different histologic criteria used in celiac disease. Clin Gastroenterol Hepatol 2007;5:838-43.
15. Montgomery EA, Voltaggio L. Biopsy Interpretation of the Gastrointestinal Tract Mucosa. Vol. 1. Philadelphia, PA: Lippincot Williams and Wilkins; 2011.
16. Fasano A, Catassi C. Current approaches to diagnosis and treatment of celiac disease: An evolving spectrum. Gastroenterology 2001;120:636-51.
17. Mubarak A, Nikkels P, Houwen R, Ten Kate F. Reproducibility of the histological diagnosis of celiac disease. Scand J Gastroenterol 2011;46:1065-73.
18. Arguelles-Grande C, Tennyson CA, Lewis SK, Green PH, Bhagat G. Variability in small bowel histopathology reporting between different pathology practice settings: Impact on the diagnosis of coeliac disease. J Clin Pathol 2012;65:242-7.
19. Taavela J, Koskinen O, Huhtala H, Lähdeaho ML, Popp A, Laurila K, *et al.* Validation of morphometric analyses of small-intestinal biopsy readouts in celiac disease. PLoS One 2013;8:e76163.
20. Iacucci M, Ghosh S. Routine duodenal biopsies to diagnose celiac disease. Can J Gastroenterol 2013;27:385.
21. Serra S, Jani PA. An approach to duodenal biopsies. J Clin Pathol 2006;59:1133-50.
22. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-44.
23. Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge, Massachusetts: The MIT Press; 2016.
24. Tomita N, Cheung YY, Hassanpour S. Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans. Comput Biol Med 2018;98:8-15.
25. Korbar B, Olofson AM, Miraflor AP, Nicka CM, Suriawinata MA, Torresani L, *et al.* Looking Under the Hood: Deep Neural Network Visualization to Interpret Whole-Slide Image Analysis Outcomes for Colorectal Polyps. CVPR Workshops; 2017. p. 69-75.
26. Byrne MF, Chapados N, Soudan F, Oertel C, Linares Pérez M, Kelly R, *et al.* Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. Gut 2019;68:94-100.
27. Korbar B, Olofson AM, Miraflor AP, Nicka CM, Suriawinata MA, Torresani L, *et al.* Deep learning for classification of colorectal polyps on whole-slide images. J Pathol Inform 2017;8:30.
28. Corral J, Hussein S, Kandel P, Bolan CW, Walla MB, Bagci U. Deep learning to diagnose intraductal papillary mucinous neoplasms (IPMN) with MRI. Gastroenterology 2018;154:S524.
29. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, *et al.* Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316:2402-10.
30. Zhou T, Han G, Li BN, Lin Z, Ciaccio EJ, Green PH, *et al.* Quantitative analysis of patients with celiac disease by video capsule endoscopy: A deep learning method. Comput Biol Med 2017;85:1-6.
31. Gadermayr M, Wimmer G, Kogler H, Vécsei A, Merhof D, Uhl A, *et al.* Automated classification of celiac disease during upper endoscopy: Status quo and quo vadis. Comput Biol Med 2018;102:221-6.
32. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Computer Vision and Pattern Recognition; 2016. p. 770-8.
33. Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. NIPS. Lake Tahoe, NV; 2012. p. 1097-105.
34. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. ICLR; 2015.
35. Russakovsky O, Deng J, Hao S, Krause J, Satheesh S, Ma S, *et al.* ImageNet large scale visual recognition challenge. IJCV 2015;115:211-52.
36. Lin TY, Maira M, Belongie S, Ludomir B, Girhisck R, Hays J, *et al.* Microsoft COCO: Common objects in context. ECCV 2014;8693:740-55.
37. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. ICCV; 2015. p. 1026-34.
38. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. ICLR; 2015.
39. Krogh A, Hertz JA. A simple weight decay can improve generalization. NIPS. Denver, CO.; 1991. p. 950-7.
40. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 1934;26:404-13.
41. Zhou B, Khosla A, Lapedriza A, Olivia A, Torralba A. Learning deep features for discriminative localization. Computer Vision and Pattern Recognition; 2016. p. 921-9.
42. LeCun Y, Cortes C. MNIST Handwritten Digit Database; 2010. Available from: https://www.yann.lecun.com/exdb/mnist. [Last accessed on 2018 Oct 10].
43. Netzer Y, Wang T, Coates A, Bissacco A, Wu B, Ng AY. Reading Digits in Natural Images with Unsupervised Feature Learning. NIPS.Granda, Spain; 2011.
44. Krasin I, Duerig T, Alldrin N, *et al.* OpenImages: a Public Dataset for Large-Scale Multi-Label and Multi-Class Image Classification; 2017. Available from: https://www.storage.googleapis.com/openimages/web/index.html. [Last accessed on 2018 Oct 22].