

# Histopathological Diagnosis System for Gastritis Using Deep Learning Algorithm

Wei Ba<sup>1#</sup>, Shuhao Wang<sup>2#</sup>, Cancheng Liu<sup>2</sup>, Yuefeng Wang<sup>2</sup>,  
Huaiyin Shi<sup>1</sup>, Zhigang Song<sup>1\*</sup>

<sup>1</sup> Department of Pathology, Chinese PLA General Hospital &  
Medical School, Beijing 100853, China

<sup>2</sup> Artificial Intelligence Lab, Thorough Images,  
Beijing 100853, China

**Key words:** artificial intelligence; deep learning; algorithm; gastritis; whole-slide pathological images

**Objective** To develop a deep learning algorithm for pathological classification of chronic gastritis and assess its performance using whole-slide images (WSIs).

**Methods** We retrospectively collected 1,250 gastric biopsy specimens (1,128 gastritis, 122 normal mucosa) from PLA General Hospital. The deep learning algorithm based on DeepLab v3 (ResNet-50) architecture was trained and validated using 1,008 WSIs and 100 WSIs, respectively. The diagnostic performance of the algorithm was tested on an independent test set of 142 WSIs, with the pathologists' consensus diagnosis as the gold standard.

**Results** The receiver operating characteristic (ROC) curves were generated for chronic superficial gastritis (CSuG), chronic active gastritis (CAcG), and chronic atrophic gastritis (CAAtG) in the test set, respectively. The areas under the ROC curves (AUCs) of the algorithm for CSuG, CAcG, and CAAtG were 0.882, 0.905 and 0.910, respectively. The sensitivity and specificity of the deep learning algorithm for the classification of CSuG, CAcG, and CAAtG were 0.790 and 1.000 (accuracy 0.880), 0.985 and 0.829 (accuracy 0.901), 0.952 and 0.992 (accuracy 0.986), respectively. The overall predicted accuracy for three different types of gastritis was 0.867. By flagging the suspicious regions identified by the algorithm in WSI, a more transparent and interpretable diagnosis can be generated.

**Conclusion** The deep learning algorithm achieved high accuracy for chronic gastritis classification using WSIs. By pre-highlighting the different gastritis regions, it might be used as an auxiliary diagnostic tool to improve the work efficiency of pathologists.

---

Received June 29, 2021; accepted August 18, 2021; published online August 31, 2021.

<sup>#</sup>Authors contributed equally to this work.

\*Corresponding author Tel: 86-13810410643; E-mail: songzhg301@139.com

Fund supported by the PLA General Hospital Medical Big Data and Artificial Intelligence Project (2019MBD-038).

## INTRODUCTION

**G**ASTRIC biopsies cover a significant part of the daily work of pathologists. Among the biopsies, chronic gastritis occupies the most of the gastric specimens. Several pathological classification systems for chronic gastritis are proposed.<sup>[1-4]</sup> In order to achieve standardization in drafting pathological reports for gastric biopsies, pathologists in China usually categorize chronic gastritis into three subgroups: chronic superficial gastritis (CSuG), chronic active gastritis (CAcG), and chronic atrophic gastritis (CAtG) in clinical workflow. With the continuous growth of the number of gastric specimens, there is a great shortage in anatomical pathologists both nationally and globally, which has created overloaded workforces for pathologists, thus effecting diagnostic accuracy of gastric biopsies.<sup>[5, 6]</sup>

Deep learning is a computer algorithm that learns by itself through a large number of examples without the need for people to set up rules. One of the deep learning algorithms, Convolutional Neural Network (CNN), has been proven to have great advantages in recognition of whole-slide images (WSIs).<sup>[7, 8]</sup> Many studies have demonstrated that CNN achieved high sensitivity and specificity in WSI classification, including lung cancer,<sup>[9]</sup> prostate cancer,<sup>[10]</sup> cutaneous melanoma,<sup>[11]</sup> bladder cancer,<sup>[10]</sup> and metastatic carcinoma of lymph node.<sup>[12, 13]</sup> However, there are limited researches about the CNN in gastritis WSI recognition. To date, there is only one study using CNN in inflammatory gastrointestinal pathology,<sup>[14]</sup> which showed promising results in interpreting gastric biopsies, including normal mucosa, *Hpylori* gastritis, and reactive gastropathy. However, the amount of data in this study is small, and chronic gastritis, which occupies most of the gastric biopsy specimens, has not been further classified. Therefore, the aim of our study is to develop a pathological classification system for chronic gastritis and test its performance using WSIs.

## MATERIALS AND METHODS

### Image data and reference standard diagnosis

The study was approved by the institutional review board of Chinese People's Liberation Army General Hospital (PLAGH) (no. S2018-163-01). The informed consents were waived because of the retrospective study and anonymization.

A total of 1,250 gastric biopsy specimens (1,128 gastritis, 122 normal mucosa) which included CSuG, CAcG, CAAtG and normal mucosa were retrospectively selected from patients who underwent biopsy for clinical care at PLAGH since October 1, 2019. Reference gold standard was established for each of 1,250 cases. Two senior specialists in pathology department independently reviewed the glass slides to confirm the diagnosis. For the slides with inconsistent diagnoses, the two specialists reviewed the case together using a multi-headed microscope to reach a consensus.

The PLAGH dataset was partitioned into three parts using random stratified sampling: (1) training dataset: 1,008 WSIs; (2) validation dataset: 100 WSIs; (3) test dataset: 142 WSIs (**Table 1**). Slides were scanned with a KF-PRO-005 (KONFOONG BIO-TECH) at a resolution of 0.238  $\mu\text{m}/\text{pixel}$ . Scanned WSIs were reviewed for image quality assessment. WSIs with defects such as missing tissue or large out-of-focus regions were rescanned.

### Image annotations

Pixel-level annotations were performed by pathologists from PLAGH on 1,250 WSIs. The annotation for CNN training was conducted using an iPad-based annotation system. In the process, different colors represent different labels. Pink, red, and purple represent CSuG, CAcG, and CAAtG, respectively. Green represents normal gastric mucosa, and brown represents poor quality and neglect area. The annotation procedure of a WSI comprised three steps: the initial labeling stage, the verification stage, and the final check stage. A slide was first randomly assigned to a pathologist. Once the labeling was finished, the slide and annotations were then passed

**Table 1.** Pathological characteristics of the whole-slide images datasets for gastritis WSIs (*n*)

| Datasets       | Normal mucosa | Chronic superficial gastritis | Chronic atrophic gastritis | Chronic active gastritis | Total |
|----------------|---------------|-------------------------------|----------------------------|--------------------------|-------|
| Training set   | 101           | 671                           | 83                         | 153                      | 1008  |
| Validation set | 10            | 29                            | 15                         | 46                       | 100   |
| Test set       | 11            | 81                            | 21                         | 29                       | 142   |

on to another pathologist for review. In the final step, a senior pathologist would check all WSIs that had passed the prior two steps. In the annotation process, false annotation would be sent back for a second-round review.

### Deep learning algorithm development

The chronic gastritis classification system used DeepLab v3 with ResNet-50 as the backbone. The deep learning algorithm was able to give pixel-level prediction. The algorithm was built with TensorFlow on an Ubuntu server. We used data parallel with four GPUs to accelerate the training process. The hyperparameters were set as follows: optimizer used ADAM (Adaptive Moment Estimation); learning rate was 0.0005; batch size was 128 (*i.e.*, 32×4 GPUs), and iteration was 400,000. The training patch size, data augmentation method, and post-processing methodologies were consistent with our gastric cancer detection algorithm.

<sup>[15]</sup> After filtered the background with Otsu's algorithm, we cropped the training slides into patches with the size of 320×320 pixels. It needs to be noted that each pixel corresponds to only one label. The training set contained 197,450, 31,708 and 19,281 patches for CSuG, CACG and CATG, respectively. A total of 126,172 patches were obtained for normal mucosa. During the training process, we used random mirroring, 90-, 180-, 270-degree rotation, brightness, contrast, hue, and saturation perturbations to enhance the robustness of the algorithm under different staining configurations.

During inference, the algorithm outputs the pixel-level probability for each gastritis type. To get the slide-level prediction, following our previous work,<sup>[15]</sup> we took the average of top 1000 pixel-level probabilities. The final heat map was displayed in the form of different colors representing various gastritis types. We used the pathologists' consensus diagnosis as the gold standard for assessing the algorithm. The inflammation type was given if the diameter of the predicted area was greater than 0.1mm (physical). We used four metrics, *i.e.*, area under the receiver operating characteristic (ROC) curve (AUC), sensitivity, specificity, and accuracy, to characterize the algorithm performance. The definitions of the metrics are:

$$\text{sensitivity} = \text{TP} / (\text{TP} + \text{FN}),$$

$$\text{specificity} = \text{TN} / (\text{TN} + \text{FP}),$$

$$\text{accuracy} = (\text{TP} + \text{TN}) / \text{Total numbers},$$

where TP, FP, TN, and FN represent true positive, false positive, true negative, and false negative, respectively.

## RESULTS

To evaluate the deep learning algorithm performance, a test set of 142 WSIs, independent of the training and validation set, was used to compare the algorithm predicted diagnosis with the gold standard diagnosis. ROC curves were generated for each type of gastritis. The deep learning algorithm achieved AUCs of 0.882, 0.910, and 0.905 for CSuG, CATG, and CACG, respectively (**Figure 1**). The sensitivity and specificity of the deep learning algorithm for the classification of CSuG was 0.790 and 1.000 (accuracy: 0.880), of CATG was 0.952 and 0.992 (accuracy: 0.986), and of CACG was 0.985 and 0.829 (accuracy: 0.901), respectively (**Table 2**). The overall predicted accuracy for three pathological subtypes of gastritis was 0.867.

A more transparent and explanatory diagnosis was provided by highlighting the gastritis types using different colors by deep learning algorithm in WSIs. Red, green, and yellow were used to represent the predicted CSuG, CATG, and CACG regions, respectively. We used probability argmax to obtain the pixel-level

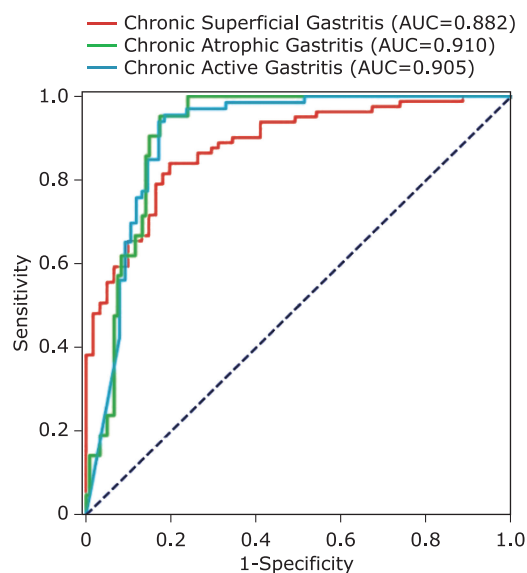


**Figure 1.** Receiver operating characteristic (ROC) curves of the deep learning algorithm for diagnosis of gastritis in the test set

**Table 2.** Efficacy metrics of the deep learning algorithm for three subtypes of gastritis

| Pathological subtypes         | Sensitivity | Specificity | Accuracy |
|-------------------------------|-------------|-------------|----------|
| Chronic superficial gastritis | 0.790       | 1.000       | 0.880    |
| Chronic atrophic gastritis    | 0.952       | 0.992       | 0.986    |
| Chronic active gastritis      | 0.985       | 0.829       | 0.901    |

predictions. **Figure 2** shows 3 cases of different gastritis in test set that were correctly diagnosed by the deep learning algorithm. **Figure 3** shows 2 cases of gastritis in test set that were incorrectly classified by the deep learning algorithm.

## DISCUSSION

In previous study, we have developed a deep learning algorithm for gastric cancer detection, and it achieved a sensitivity near 100% and an average specificity of 80.6% in a real-world testing dataset

(3,212 WSIs).<sup>[15]</sup> In this current study, we developed a deep learning algorithm for pathological classification of chronic gastritis. Our results demonstrated that the performance of the deep learning algorithm achieved accuracies of 0.880, 0.986, and 0.901 for CSuG, CAtG, and CACG, respectively. Because of the lack of pathologists and the increasing incidence of cancer, pathologists currently have to examine a large number of slides to make a final diagnosis. The biggest benefit of flagging regions of interest is that it reveals insights into network decisions. As shown in **Figure 2**, the majority of chronic gastritis subtypes in WSIs were

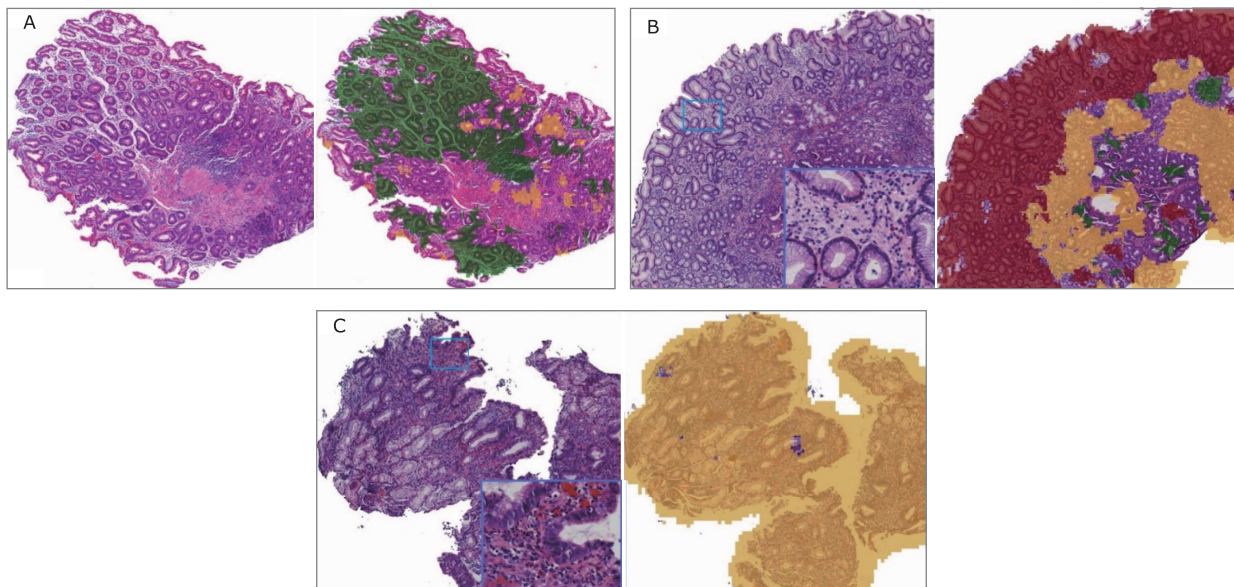


**Figure 2.** Examples of chronic atrophic gastritis (CAtG), chronic superficial gastritis (CSuG) and chronic active gastritis (CACG) in test data set that were correctly diagnosed by the deep learning algorithm

(A) CAtG. (B) CSuG. (C) CACG. Images on the left represent typical histological findings for the disease diagnosis. The green, red, and yellow in the images on the right represent the predicted CAtG, CSuG, and CACG regions, respectively. The boxes in the lower right corner of (B) and (C) images represent the local magnification of the corresponding pane in the image, which show the infiltrating inflammatory cells of different gastritis.

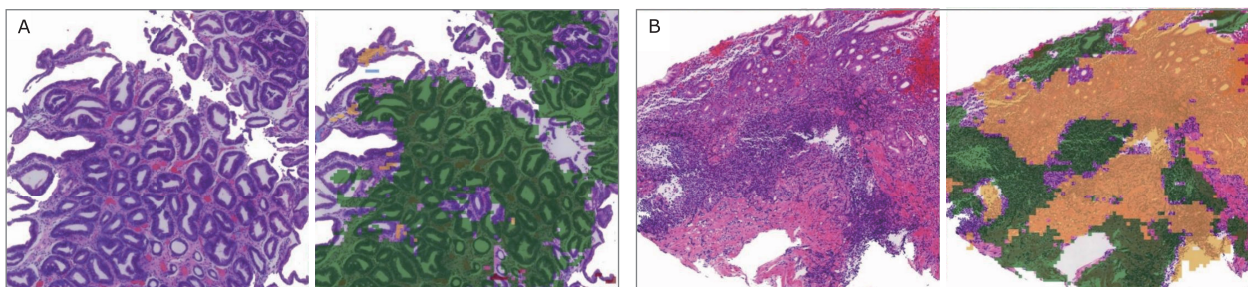


**Figure 3.** Examples of chronic superficial gastritis (CSuG) and chronic atrophic gastritis (CAtG) that were incorrectly diagnosed by the deep learning algorithm

(A) CSuG was misdiagnosed as CAtG. (B) CAtG was misdiagnosed as chronic active gastritis (CACG). Images on the right are the outputs of the deep learning algorithm. The green and yellow colors represent the predicted CAtG and CACG regions by the algorithm.

correctly recognized, which confirmed that the neural network we used could accurately distinguish features to make a decision. Therefore, this technique might be employable as a supplemental screening tool by automatically pre-screening and highlighting the interested regions prior to review. Compared with the research of Martin and his colleagues, our algorithm has larger volume of training data and more precise annotation of the images, which may be the reasons for the better performance.<sup>[14]</sup>

We performed a thorough analysis on the predicted results by deep learning algorithm to further improve the pathologists' confidence for the artificial intelligence system. As shown in **Figure 3A**, CSuG was misdiagnosed as CATG due to some degree of morphological overlap between these two kinds of gastritis in some areas. For CATG, the number of glands was significantly reduced within mucosal layer. As shown in **Figure 3B**, the deep learning algorithm mistakenly predicted the smooth muscle layer as the area of CATG, probably because there is no gland structure in this region, which demonstrated that the algorithm did not fully learn the differences between CATG and normal mucosa. Although in the training process of deep learning algorithm, we have added normal gastric mucosa WSIs, expanding the sample size is needed in future study to avoid this mistake. Furthermore, in the prediction process, the algorithm also mistakenly classified the normal mucosa as CSuG, which may attribute to that normal gastric mucosa may be inevitably labeled as CSuG in annotation process, because practically normal gastric mucosa is always mixed with CSuG regions in a specimen. Other reasons for false results include poor image quality, such as section folds, knife marks, and overstaining, as have demonstrated in previous studies.<sup>[7, 11, 16]</sup>

This study had several limitations. The pathological subtypes of gastritis were complex. Our study only included the most common chronic gastritis encountered in routine pathological workflow. Secondly, differences in the inherent difficulty of the test set will directly affect the diagnostic performance of the deep learning algorithm. In our study, the training and testing data set come from one medical center, for which the algorithm is likely to perform differently on external test set. Future studies using WSIs from multiple centers in training and testing datasets are needed to make the system universally useful. In addition, prospective studies are needed to confirm our results with larger and more diverse datasets, and to address pathologists' acceptance

of this new technique.

In summary, our research shows that the well-trained deep learning algorithm has a high degree of accuracy when using WSI to classify chronic gastritis. Therefore, the deep learning algorithm can not only be used as a supplementary diagnostic tool to highlight different areas of gastritis in advance for review, but also as an analog to a second opinion from fellow pathologists.

### **Conflict of interests**

*Shuhao Wang is the co-founder and chief technology officer (CTO) of the Thorough Images. Cancheng Liu and Yuefeng Wang are algorithm researchers of the Thorough Images. All remaining authors have declared no conflict of interests.*

### **REFERENCES**

1. Correa P, Yardley JH. Grading and classification of chronic gastritis: one American response to the Sydney system. *Gastroenterology* 1992; 102(1):355-9. doi: 10.1016/0016-5085(92)91820-t.
2. Price AB, Misiewicz JJ. Grading and classification of chronic gastritis: the response of the working party. *Gastroenterology* 1992; 103(3):1116-7. doi: 10.1016/0016-5085(92)90066-8.
3. Dixon MF, Genta RM, Yardley JH, et al. Classification and grading of gastritis. The updated Sydney System. International Workshop on the Histopathology of Gastritis, Houston 1994. *Am J Surg Pathol* 1996; 20(10):1161-81. doi: 10.1097/00000478-199610000-00001.
4. Stolte M, Meining A. The updated Sydney system: classification and grading of gastritis as the basis of diagnosis and treatment. *Can J Gastroenterol* 2001; 15(9):591-8. doi: 10.1155/2001/367832.
5. Metter DM, Colgan TJ, Leung ST, et al. Trends in the US and Canadian Pathologist Workforces From 2007 to 2017. *JAMA Netw Open* 2019; 2(5):e194337. doi: 10.1001/jamanetworkopen.2019.4337.
6. Retamero JA, Aneiros-Fernandez J, Del Moral RG. Complete digital pathology for routine histopathology diagnosis in a multicenter hospital network. *Arch Pathol Lab Med* 2020; 144(2):221-8. doi:10.5858/arpa.2018-0541-OA.
7. Cruz-Roa A, Gilmore H, Basavanahally A, et al. Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for

- quantifying tumor extent. *Sci Rep* 2017; 7:46450. doi: 10.1038/srep46450.
8. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 2017; 318(2):2199-210. doi: 10.1001/jama.2017.14585.
  9. Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nature Medicine* 2018; 24(10):1559. doi: 10.1038/s41591-018-0177-5.
  10. Khosravi P, Kazemi E, Imielinski M, et al. Deep convolutional neural networks enable discrimination of heterogeneous digital pathology images. *EBioMedicine* 2018; 27:317-28. doi: 10.1016/j.ebiom.2017.12.026.
  11. Ba W, Wang R, Yin G, et al. Diagnostic assessment of deep learning for melanocytic lesions using whole-slide pathological images. *Transl Onc* 2021; 6(1):475. doi: 10.1016/j.tranon.2021.101161.
  12. Golden JA. Deep learning algorithms for detection of lymph node metastases from breast cancer: helping artificial intelligence be seen. *JAMA* 2017; 318(22):2184-6. doi: 10.1001/jama.2017.14580.
  13. Pan Y, Sun Z, Wang W, et al. Automatic detection of squamous cell carcinoma metastasis in esophageal lymph nodes using semantic segmentation. *Clin Transl Med* 2020; 10(3):e129. doi: 10.1002/ctm2.129.
  14. Martin DR, Hanson JA, Gullapalli RR, et al. A deep learning convolutional neural network can recognize common patterns of injury in gastric pathology. *Arch Pathol Lab Med* 2020; 144(3):370-8. doi: 10.5858/arpa.2019-0004-OA.
  15. Song Z, Zou S, Zhou W, et al. Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning. *Nat Commun* 2020; 11(1):4294. doi: 10.1038/s41467-020-18147-8.
  16. Arvaniti E, Fricker KS, Moret M, et al. Automated Gleason grading of prostate cancer tissue microarrays via deep learning. *Sci Rep* 2018; 8(1):12054. doi: 10.1038/s41598-018-30535-1.
  17. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019; 25(1):44-56. doi: 10.1038/s41591-018-0300-7.
  18. Tschandl P. Problems and potentials of automated object detection for skin cancer recognition. *JAMA Dermatol* 2020; 156(1):23-4. doi: 10.1001/jamadermatol.2019.3360.