Gastroenterologist-level detection of gastric precursor lesions and neoplasia with a deep convolutional neural network

Short title: Detection of gastric lesions with neural network

Lei Chen[1], Shengtao Zhu[1], Wenjie Chen[2], Li Min[1], Yu Zhao[1], Fengtong Du[2], Shanshan Wu[3], Shuilong Guo [1], Jie Xing[1], Zheng Zhang[1], Hongtao Li[4], Ming Ji[1], Peng Li[1], Lihong Cao[2]*, Shutian Zhang[1]*

[1]Department of Gastroenterology, Beijing Friendship Hospital, Capital Medical University, National Clinical Research Center for Digestive Disease, Beijing Digestive Disease Center, Beijing Key Laboratory for Precancerous Lesion of Digestive Disease. No.95, Yong'an Rd, Xicheng District, Beijing,100050, China

[2]Neuroscience and Intelligent Media Institute, Communication University of China, Intelligent Media Center, Beijing Institute of Collaborative Innovation. Building No.42, No.1 East DingFuZhuang St, Chaoyang District, Beijing, 100024, China.

[3]Department of Biostatistics, Beijing Friendship Hospital, Capital Medical University, National Clinical Research Center for Digestive Disease. No.95, Yong'an Rd, Xicheng District, Beijing,100050, China.

[4]Beijing Hotwire Medical Tech Development Co., Ltd. Building No.1, Scientific Incubator Center, No.1 ScienceCityNuclear Street, Fengtai District, Beijing, 100070, China.

Lei Chen, Shengtao Zhu and Wenjie Chen contributed equally to this work.

*Corresponding author: Prof. Shutian Zhang, Department of Gastroenterology, Beijing Friendship Hospital, Capital Medical University, National Clinical Research Center for Digestive Disease, Beijing Digestive Disease Center, Beijing Key Laboratory for

Precancerous Lesion of Digestive Disease, Beijing, 100050, China. E-mail: zhangshutian@ccmu.edu.cn

Prof. Lihong Cao, Neuroscience and Intelligent Media Institute, Communication University of China, Intelligent Media Center, Beijing Institute of Collaborative Innovation. Beijing, 100024, China. E-mail: lihong.cao@cuc.edu.cn.

ABSTRACT

Gastric precursor lesions and neoplasia with very delicate changes in the gastric mucosa could be easily missed or misdiagnosed under endoscopy. Here we developed an automatic real-time pattern recognition tool based on convolutional neural networks (CNNs) algorithm to help endoscopists in detection of chronic atrophic gastritis (CAG) and gastric cancer (GC) lesions. The five-convolution-layer ZF model and thirteen-convolution-layer VGG16 model were combined in our neural network A total of 10,014 CAG and 3,724 GC annotated images were used in the network training. Another independent set consisted of 50 CAG, 50 GC and 100 negative controls images were used to evaluate the performance of the final network. In CAG detection, the performance of our model was much better than the average performance of the 77 endoscopists in sensitivity, specificity and accuracy (95% versus 74%, 86% versus 82%, 90% versus 78%, respectively). In GC detection, the performance of our model achieved a slightly higher sensitivity (90% versus 87%), but a lower specificity (50% versus 74%) and accuracy (70% versus 80%) than the average performance of the 89 endoscopists. In conclusion, we provided a CNN based computational tool to improve the detection of CAG and GC under endoscopy, and simplify diagnostic procedures.

Keywords: Gastric precursor lesion, Gastric cancer; Deep learning, Convolutional neural network, Computer-aided endoscopic detection support system

INTRODUCTION

Gastric cancer (GC) is reported as the fifth most commonly diagnosed malignancy in the world with about one million new cases occurred in 2012 (951,000 cases, 6.8% of the total).[1] More than 70% of GC cases (677,000 cases) occurred in developing countries, and half occurred in Eastern Asia (mainly in China). Additionally, GC is the third leading cause of cancer death in both sexes worldwide (723,000 deaths, 8.8% of the total).[1] However, the prognosis of GC varies a lot among different stages. The 5-year survival rate of early gastric cancer (EGC) almost exceeds 90%, whereas less than 20% advanced GC patients can survive for more than 5 years.[2,3] Therefore, early detection and regular surveillance in the high-risk population are probably the most effective strategies to improve the survival rate of GC. A multistep progression which is predominantly triggered by H. pylori infection and followed by chronic gastritis, atrophic gastritis (AG), intestinal metaplasia (IM) and finally intestinal-type GC, has been widely accepted.[4-7] Particularly, AG is considered as a necessary transitional step in gastric carcinogenesis,[5,8] which is characterized by chronic inflammatory processes of gastric mucosa that leads to the loss of glandular structure and a reduction of gastric secretory function.[8] One cohort study indicated that the annual incidence rates of GC in patients with AG is 4.5 times of that in the general population.[9] Thus, accurate detection of AG along with regular surveillance and subsequent management could be very helpful to control GC in an early stage.

However, many premalignant lesions and EGCs with very delicate changes in the gastric mucosa demand extremely careful observation and inspectional skills, and such lesions (especially superficial flat ones) may be easily missed or misdiagnosed by conventional white light imaging (WLI) endoscopy. Recently, a retrospective cohort study in England reported an endoscopy miss rate of approximately 8.3% in 2,727 patients with GC.[10] Various advanced endoscopic modalities have been developed to improve the detection rate and diagnostic accuracy, such as high definition endoscopy (HDE), narrow band imaging (NBI),[11] magnifying endoscopy, chromoendoscopy and etc.[12] Nevertheless, the advanced endoscopic techniques are very expensive, and additional operation trainings are also required, making it impossible to be widely

utilized in primary health centers. A readily accessible, cost-effective and comparatively reliable diagnostic approach for detecting premalignant lesions and ECG was strongly needed.

Due to the development of big data, deep learning algorithms have become the research focus of artificial intelligence.[13] Convolutional neural networks (CNNs), known as the most successful deep learning strategy applied to image classification,[14] have brought about a revolution in computer vision.[13] CNNs can extract a set of transformations from inputted data automatically and avoid manual design of specific feature detectors.[14] Using CNNs to analysis biomedical image has become more and more popular in many clinical scenarios, such as classification of histologic and histopathologic images,[15,16] diagnosis of Alzheimer disease,[17-20] differentiating breast lesions[21] and recognition of skin cancers.[22] However, few works have explored to address the automatic diagnosis of gastric premalignant lesions and neoplasia.

In this study, we constructed two independent gastrointestinal (GI) image datasets, and fine-tuned two types of deep learning models named ZF[23] and VGG16[24] based on Faster R-CNN (Faster region-based convolutional neural network)[25] to identify CAG and GC lesions. The results were compared with GI doctors with different seniority to evaluate the performance of those models.

MATERIALS AND METHODS

Patients Information and Study Design

We conducted a single-center, retrospective diagnostic study and it was performed after the protocol approved by the Institutional Review Board and ethics committees of Beijing Friendship Hospital, Capital Medical University. We reviewed our endoscopic database to identify all patients with diagnosis of CAG and GC (in both early and advanced stage) from January 1st in 2013 to June 10th in 2017. Informed consent was not required because only de-identified patient data were obtained.

Data acquisition and processing

Distinct endoscopic images and relevant medical records of the patients were finally extracted when they fulfill either of the following criteria: (1) Diagnosis of CAG should

be proved by histological classification of the Updated Sydney System after applying its gastric biopsy sampling protocol.[26] (2) Diagnosis of GC must be endoscopically confirmed together by any two out of the nine specific GI experts certified by Chinese Gastroenterological Endoscopic Society. The diagnostic criteria were mainly based on personal experience according to morphological features of lesions with or without pathological results.

Exclusion criteria were: (1) The biopsy sites of CAG did not strictly adhere to the standard endoscopic biopsy sampling procedure mentioned in the Updated Sydney System or lesions in the endoscopic images were difficult to be identified; (2) The endoscopic images showed indistinct involvement of GC; (3) The patients who have comorbidity of malignancy of other systems; (4) The endoscopic images appear to be unclear and/or the shooting angles do not reach our requirements.

For both CAG and GC detection tasks, we established a training dataset and a testing dataset with non-overlapping requested images (Figure 1A). The testing datasets which consisted of equivalent positive and negative samples were prepared for model validation. As for CAG, we used images of chronic superficial gastritis (CSG) as its negative samples. And with respect to negative samples for GC, we enrolled images that fulfilled at least one of the following diagnosis: benign gastric ulcer and polyps, gastric stromal tumors and gastric heterotopia pancreas (all proved by histopathological results) (Supplementary Table S1). All images were de-identified immediately.

Data annotation

Six experienced endoscopists were recruited to annotate images in the training datasets with bounding boxes. The boxes were supposed to be drawn in the exact biopsy sites according to both endoscopic descriptions and histological results. Each image was annotated by two endoscopists back-to-back (See Supplementary Materials and Methods). If the bounding boxes annotated by different endoscopists did not differ a lot from each other, the intersection area of both boxes were adopted as the final annotation. If the boxes were poorly overlapped (intersection area < 50% of the total area), such images would be picked out and would be discussed together by all the 6 endoscopists mentioned above. As a result, 364 images of CAG and 92 images of GC were

afterwards discussed together by all the doctors.

Modified Faster R-CNN

An overview of the structure of Faster R-CNN is shown in Figure 1B. In training process, we compared a five-convolution-layer ZF model[23] with a thirteen-convolution-layer VGG16 model[24] based on Faster R-CNN. We firstly pre-trained both the ZF and the VGG16 model with ImageNet dataset and randomly initialized all new layers by drawing weights from a zero-mean Gaussian distribution with standard deviation of 0.01. We fine-tuned the whole layers of region proposal network (RPN) and the final bounding box regression and classification layers at the same time so that the standard four-step alternating algorithm of Faster R-CNN training process could be modified to an end-to-end one. The end-to-end process was more efficient for speeding up the training procedure and getting higher detective quality. We also applied the 'image-centric' sampling strategy. A number of mini-batches were generated from a single image that contained many positive and negative anchors. We randomly sampled 256 anchors in each image to compute the loss function of a mini-batch where the ratio of positive and negative anchors was 1:1. If there were fewer than 128 positive samples in an image, we padded the mini-batches as negative ones. Learning rate of 0.001 was used for 50k mini-batches and 0.0001 for the next ones.

The readily processed network was afterwards used to finish the CAG and GC detection test with the threshold of classification score as 0.85. It was thought to be negative if none of suspicious lesion was detected in a single testing image. Four statistical parameters named TP FN FP TN were calculated (TP, True Positive; FN, False Negative; FP, False Positive; TN, True Negative). The final accuracy of different datasets was revealed for identifying the performance of Faster R-CNN.

Performance Evaluation of GI Doctors

Every GI doctor involved in the validation tasks was assigned to the same testing dataset as the computational models were. Their performance was firstly evaluated by overall sensitivity, specificity and accuracy against histopathological diagnosis. Besides, all the doctors in each test were stratified into four levels (from Level I to Level IV) by years of endoscopic operation (Level I, < 5 years; Level II, 5~10 years; Level III, 10~15

years; Level IV, ≥ 15 years). We further estimated the average diagnostic reliability in every one of the four levels.

Statistical Analysis

We evaluated performance of networks and GI doctors by calculating sensitivity, specificity, accuracy, positive predictive value (PPV), and negative predictive value (NPV). Inter-observer agreement of GI doctors was evaluated by Fleiss kappa measurement (more than 2 observers) with 95% confidence interval.[27] Interpretation of kappa values was done according to Landis and Koch.[28] Comparisons between the best computational model and GI doctors in sensitivity, specificity, and accuracy were analyzed by Pearson's chi-squared test. *P*-value < 0.05 was considered statistically significant.

RESULTS

Description of Datasets

Totally, 10064 annotated images of CAG with definite histopathological results were obtained, among which 50 images were randomly set aside for testing. Another 50 images of CSG were also included as negative samples in the testing dataset. We then input the remaining 10014 CAG images into Faster R-CNN network for machine learning. Examples of annotated images of CAG are demonstrated in Figure 2A.

Similarly, 3774 annotated images of GC concurrently endoscopically diagnosed by GI experts were collected. To be specific, 1540 images were diagnosed as EGC, among which 462 images had definite pathological verification. We therefore randomly extracted 50 out of the 3774 images and combined them with additional 50 images of non-cancerous lesions mentioned above to set up the testing dataset. The remaining 3724 GC images were put into training. Examples of annotated images of GC are demonstrated in Figure 2B.

Performance of Faster R-CNN

For both CAG and GC detection, the ZF and the VGG16 models were trained meanwhile built upon a Faster RCNN architecture. Performance of these models were evaluated upon the testing dataset. We chose the model with the best accuracy to

represent the final performance and to compare with GI doctors (Table 1).

A highest accuracy of 90% (90/100) was achieved by adapting the VGG16 model with iteration of 50000 and threshold of 0.85 in CAG detection task (Supplementary Table S2A, sensitivity, 94.0%; specificity, 86.0%; PPV, 87.0%; and NPV, 93.5%). As for GC detection, the ZF model with iteration of 100000 and threshold of 0.85 achieved the optimal accuracy of 70% (70/100) (Supplementary Table S2B, sensitivity, 90.0%; specificity, 50.0%; PPV, 64.3%; and NPV, 83.3%).

Performance of GI Doctors

There were 77 and 89 GI doctors with different seniority taking the CAG and GC detection test respectively. Their baseline characteristics were presented in Supplementary Table S3.

For CAG detection, the sensitivity, specificity and accuracy of the all 77 GI doctors with different seniority respectively range within 16%~100% (median 78%, average 74%), 0%~94% (median 88%, average 82%), and 21%~87% (median 81%, average 78%). The sensitivity of Level I to Level IV was separately 61.4%, 72.8%, 82.2% and 79.8% while the specificity was respectively 78.2%, 73.8%, 81.4% and 85.4%. The accuracy of them was respectively 69.8%, 73.3%, 81.1% and 82.6%. Smooth rising trends were observed in all of the three statistical parameters, despite a slight decrease of specificity in Level II compared with Level I.

For GC detection, the sensitivity, specificity and accuracy respectively range within 48%~100% (median 88%, average 87%), 0%~62% (median 78%, average74%), and 35%~73% (median 82%, average 80%). The sensitivity of Level I to Level IV was respectively 84.6%, 83.6%, 89.6% and 87.4%. The specificity ranged from 68.8%, 70.0%, 73.2% to 79.4%, while the accuracy was 76.7%, 76.8%, 81.4% and 83.4% respectively. The above three parameters completely presented as incremental diagrams in the bar chart from Level I to Level IV in spite of slight reduction of sensitivity in Level II and Level IV.

The inter-observer agreement of doctors in different levels regarding diagnosis of CAG and GC is listed in Supplementary Table S4. The best agreement was all obtained by doctors in Level IV.

Comparison of Performance between Faster R-CNN and GI Doctors

Compared with average performance of the 77 doctors, performance of the best model is much better in sensitivity, specificity and accuracy (95% versus 74%, 86% versus 82%, 90% versus 78%, respectively).

After classifying all the GI doctors based on seniority, an overview of performance the optimal model and the 77 doctors is illustrated in Supplementary Figure 1A. Both TP and TN of the optimal network successfully exceed those of the level IV doctors, indicating that sensitivity and specificity as well as accuracy of Faster R-CNN have already reached expert-level.

Statistical differences between the network and different-level doctors are observed in the sensitivity (all levels, $Ps < 0.05$) and accuracy (all levels, $Ps < 0.05$, except Level III, $P = 0.103$). However, there is no significant difference in specificity (all levels, $Ps > 0.05$, except Level II, $P = 0.034$, shown in Figure 3A).

Except a slightly elevated sensitivity (90% versus 87%), the specificity and accuracy of our network are significantly lower than the average performance of the doctors (50% versus 74%, 70% versus 80%, respectively).

An overview of performance between Faster R-CNN and the 89 doctors classified to different levels is demonstrated in Supplementary Figure 1B. The sensitivity of our network is equal to GI experts (Level IV doctors) while the specificity is much lower.

There is no significant difference between the network and doctors of different levels in sensitivity (all levels, $Ps > 0.05$) and accuracy (all levels, $Ps > 0.05$, except Level IV, $P = 0.03$). However, significant difference in specificity was observed (all levels, $Ps < 0.05$). The specificity of our network is much lower than doctors in all different levels (Figure 3B).

DISCUSSION

In this study, we have provided an automatic real-time lesion detector focusing on endoscopic diagnosis of CAG and GC based on deep convolutional neural network algorithm. The performance of our model was also evaluated and compared to board-certified GI doctors of different seniorities.

For CAG detection, the best model achieved a gastroenterologist-level performance.

The sensitivity, specificity and accuracy (95%, 86%, 90%, respectively) of this model all exceed those of the level IV (the highest seniority) doctors. For GC detection, the best model achieved superior sensitivity (90% versus 87%) but inferior specificity and accuracy (50% versus 74%, 70% versus 80%, respectively) compared with average performance of all the 89 doctors. These results suggested that the network made positive diagnosis as much as possible and consequently aggravates misdiagnosis of non-cancer lesions. Considering that it is a standard procedure to take subsequent biopsies for making definite diagnosis before treatment, the over-diagnose of our model to ensure a high sensitivity would be acceptable for endoscopists.

Some GCs showed very slight change of mucosa, especially some superficial lesions, which brought challenges to endoscopists. For these lesions, our network may server as a useful complement to human eyes. Some images of unnoticeable GCs correctly detected by our network are shown in Supplementary Figure 2. We also extracted the negative cases misdiagnosed as GCs by the network (Supplementary Figure 3). Because of high similarity of morphology with GCs, inputting a certain number of such images for training would be helpful to reduce the false positive rate. Besides, the inter-observer agreement is unsatisfactory even within doctors of Level IV (Kappa, 0.584). In contrast to obvious diagnostic variations of doctors, our network is stable, uniform and repeatable.

Although several studies have been reported for applying computer-aided system to classifying colonic[29,30] and pancreatic lesions[31-34] few work contributes to gastric premalignant lesions and neoplasia. Besides, most methods mentioned above were focused on differentiating, not detecting. Actually, real-time detecting need much stronger capacity of pattern recognition than differentiating, which may be far beyond the ability of traditional machine learning models based on k-NearestNeighbor (kNN) and Support Vector Machine (SVM). In this study, we constructed several models based on CNNs, the most powerful deep learning algorithms at present.[13] To achieve superior stability, reliability and accuracy, we trained a modified Faster R-CNN with a database of 13738 endoscopic images (including 10014 images for CAG detection and 3724 images for GC detection), which is hundreds of times larger than that of previous

studies.

Traditional algorithms usually demand for manual extraction of domain-specific visual features, followed by further ability of classifying. Therefore, their application may be greatly restricted for the disability of automatic discovery and location of lesions. Additionally, since most of the precancerous lesions and early neoplasia are presented as subtle alteration of morphology and color of mucosa, they are quite difficult to be precisely detected only with WLI during examination even by some experts. Thus, we designed and implemented an automatic real-time lesion detector for CAG and GC, taking advantage of independent learning from little pre-processing sources by Faster R-CNN.

CAG lesions often present as diffused mucosal changes, which makes it much more difficult to delineate the outline of the lesion than other diseases such as GC. Bounding boxes for annotating positive samples are placed in biopsy sites manually which usually only cover the most severe area of lesions. However, in training process, we further generated a group of candidate bounding boxes widely distributed in the whole image based on the 'anchor rules' mentioned in the previous study.[25] We then labelled them as positive/negative according to Intersection over Union (IoU) between the ground-truth boxes and each of them. After that, they were randomly selected and inputted for training. Subsequently, some candidate bounding boxes labeled negative may actually include sporadic atrophic lesions, leading to a small amount of noise in negative samples. Based on atypical characteristics and small percentage of such noise, our network can adjust and converge itself to a relatively satisfactory condition after a long period of self-learning (Supplementary Video 1). Expert-level performance in CAG detection also proves strong ability to correctly distinguish positive samples from negative ones.

In GC detection, we enrolled both early and advanced GC diagnosed endoscopically together by two of the nine certified GI experts with or without pathological results in order to make full use of all data, which means the training set may include some over-diagnosed images, resulting in enhanced sensitivity and underestimated specificity and accuracy of our network with a standard of histopathology in the test. A larger number

of GC images with definite histopathological results would be greatly needed in training dataset to obtain a better network. Additionally, all images of EGC would be picked out to improve the detection rate of the most controversial but fatal cancerous lesions.

The primary motivation for designing a real-time computer-aided lesion detector in endoscopy is to assist young GI doctors in discovering and precisely locating gastric precursor lesions and neoplasia, especially EGCs. We further hope to diminish the false dismissal rate and misdiagnosis rate of EGC, as well as helping to direct specific biopsy sites. In real clinical environment, the readily processed network will be converted to a software, and integrated with endoscopic operating system. There is no need in altering endoscopic examination protocol nor hardware to use our system. All we need is WLI rather than any other advanced endoscopic equipment. After a long-term follow-up observation, the preconceived reduction in false dismissal rate and misdiagnosis rate of CAG, EGC and AGC (advanced GC) would be statistically calculated. The difference of overall medical cost per patient between the novel endoscopic diagnosis procedure and the conventional one will also be analyzed. Moreover, the system could also be used as an educational tool, speeding up the learning curve of endoscopic beginners.

For CAG detection, our network outperforms GI experts in sensitivity, specificity and accuracy. For GC detection, our network has a superior sensitivity but inferior specificity and accuracy than GI experts. In conclusion, we provided a deep learning based computational tool to improve the detection rate of CAG and GC, simplify diagnostic procedures and target following biopsy.

Acknowledgements

We would like to thank Dr. Xiujing Sun, Juan Liu, Zheng Zhang, Rui Cheng, Xun Yang, Yue Jiao, Changqin Xu, Haiying Zhao, Fandong Meng, Wenyan Li, Fujing Lv and Ye Zong for their help in dataset preparation and endoscopic image interpretation.

Funding

We appreciate the financial support received from Beijing Municipal Science and Technology Commission (Z171100004517009) and the study sponsor had no role in

this manuscript.

Conflict of Interest

All the authors declare that there is no interests and involvements which have a bearing on the paper.

Ethics Approval

This study protocol was approved by the Ethics Committee of the Beijing Friendship Hospital, Capital Medical University. Serial number: 2020-P2-090-01.

REFERENCES

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, *et al*. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 2015; 136(5).

2. Yuasa, N. and Y. Nimura. Survival after surgical treatment of early gastric cancer, surgical techniques, and long-term survival. Langenbecks Arch Surg 2005; 390(4): 286-293.

3. Park JM, Ryu WS, Kim JH, Park SS, Kim SJ, Kim CS, *et al*. Prognostic factors for advanced gastric cancer: stage-stratified analysis of patients who underwent curative resection. Cancer Res Treat 2006; 38(1):13-8.

4. Kato S, Matsukura N, Tsukada K, Matsuda N, Mizoshita T, Tsukamoto T, *et al*. Helicobacter pylori infection-negative gastric cancer in Japanese hospital patients: incidence and pathological characteristics. Cancer Sci 2007; 98(6):790-4.

5. Kim N, Park RY, Cho SI, Lim SH, Lee KH, Lee W, *et al*. Helicobacter pylori infection and development of gastric cancer in Korea: long-term follow-up. J Clin Gastroenterol 2008; 42(5):448-54.

6. Siurala, M., K. Varis, and M. Wiljasalo, Studies of patients with atrophic gastritis: a 10-15-year follow-up. Scand J Gastroenterol 1966; 1(1): 40-48.

7. Ohata H, Kitauchi S, Yoshimura N, Mugitani K, Iwane M, Nakamura H, *et al*. Progression of chronic atrophic gastritis associated with Helicobacter pylori infection increases risk of gastric cancer. Int J Cancer 2004; 109(1):138-43.

8. De Vries, A.C., J. Haringsma, and E. Kuipers. The detection, surveillance and treatment of premalignant gastric lesions related to Helicobacter pylori infection. Helicobacter 2007; 12(1):1-15.

9. de Vries AC, van Grieken NC, Looman CW, Casparie MK, de Vries E, Meijer GA, *et al*. Gastric cancer risk in patients with premalignant gastric lesions: a nationwide cohort study in the Netherlands. Gastroenterology 2008; 134(4):945-52.

10. Chadwick G, Groene O, Riley S, Hardwick R, Crosby T, Hoare J, *et al*. Gastric Cancers Missed During Endoscopy in England. Clin Gastroenterol Hepatol 2015; 13(7):1264-1270.

11. Kaise M. Advanced endoscopic imaging for early gastric cancer. Best Pract Res Clin Gastroenterol 2015; 29(4):575-87.

12. Moon HS. Improving the Endoscopic Detection Rate in Patients with Early Gastric Cancer. Clin Endosc 2015; 48(4):291-6.

13. Deng, J. Imagenet: A large-scale hierarchical image database. in Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. 2009. IEEE.

14. Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw 2015;61:85-117.

15. Arevalo JA. González. Hybrid image representation learning model with invariant features for basal cell carcinoma detection. in Proc. SPIE. 2013.

16. Cruz-Roa AA, Arevalo Ovalle JE, Madabhushi A, González Osorio FA. A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection. Med Image Comput Comput Assist Interv 2013; 16(2):403-10.

17. Suk HI, Shen D. Deep learning-based feature representation for AD/MCI classification. Med Image Comput Comput Assist Interv 2013; 16(2):583-90.

18. Suk HI, Lee SW, Shen D. Alzheimer's Disease Neuroimaging Initiative. Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. Brain Struct Funct 2015; 220(2):841-59.

19. Suk HI, Lee SW, Shen D. Alzheimer's Disease Neuroimaging Initiative. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. Neuroimage 2014; 101:569-82.

20. Li F. Robust deep learning for improved classification of AD/MCI patients. in International Workshop on Machine Learning in Medical Imaging. Springer. 2014.

21. Jamieson AR, K Drukker, ML Giger. Breast image feature learning with adaptive deconvolutional networks. SPIE Medical Imaging Strony 2012; 831506-831506.

22. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017; 542(7639):115-118.

23. Zeiler MD, R Fergus. Visualizing and understanding convolutional networks. in

European conference on computer vision. Springer. 2014.

24. Simonyan K, A Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

25. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans Pattern Anal Mach Intell 2017; 39(6):1137-1149.

26. Dixon MF, Genta RM, Yardley JH, Correa P. 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.

27. Fleiss JL, JC Nee, JR Landis. Large sample variance of kappa in the case of different sets of raters. Psychological Bulletin 1979; 86(5):974.

28. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33(1):159-74.

29. Takemura Y, Yoshida S, Tanaka S, Kawase R, Onji K, Oka S, *et al*. Computer-aided system for predicting the histology of colorectal tumors by using narrow-band imaging magnifying colonoscopy (with video). Gastrointest Endosc 2012; 75(1):179-85.

30. Kominami Y, Yoshida S, Tanaka S, Sanomura Y, Hirakawa T, Raytchev B, *et al*. Computer-aided diagnosis of colorectal polyp histology by using a real-time image recognition system and narrow-band imaging magnifying colonoscopy. Gastrointest Endosc 2016;83(3):643-9.

31. Zhu J, Wang L, Chu Y, Hou X, Xing L, Kong F, *et al*. A new descriptor for computer-aided diagnosis of EUS imaging to distinguish autoimmune pancreatitis from chronic pancreatitis. Gastrointest Endosc 2015; 82(5):831-836.

32. Das A, Nguyen CC, Li F, Li B. Digital image analysis of EUS images accurately differentiates pancreatic cancer from chronic pancreatitis and normal tissue. Gastrointest Endosc 2008; 67(6):861-7.

33. Zhang MM, Yang H, Jin ZD, Yu JG, Cai ZY, Li ZS. Differential diagnosis of pancreatic cancer from normal tissue with digital imaging processing and pattern recognition based on a support vector machine of EUS images. Gastrointest Endosc 2010; 72(5):978-85.

34. Zhu M, Xu C, Yu J, Wu Y, Li C, Zhang M, *et al*. Differentiation of pancreatic cancer and chronic pancreatitis using computer-aided diagnosis of endoscopic ultrasound (EUS) images: a diagnostic test. PLoS One 2013; 8(5):e63820.

Figures Legends

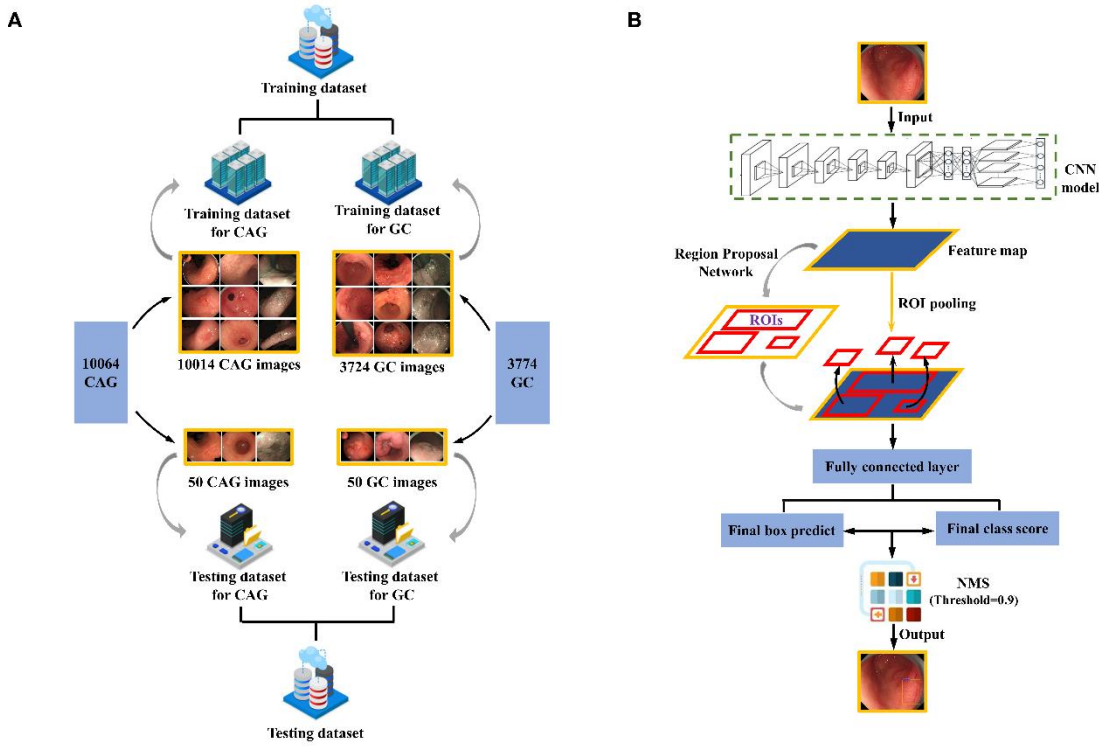

Figure 1. Schematic illustration of data composition and processing and an overview of Faster R-CNN structure.

(A) Composition of training and testing dataset for both CAG and GC and post-processing before training. 10064 images of CAG were extracted from our endoscopic database, among which 50 images were randomly set aside. The other 10014 images consisted of the training dataset for CAG. The testing dataset contained another 50 images of CSG randomly selected as negative samples and 50 images of CAG pick-out before. Similarly, 3774 images of GC were extracted and we picked out 50 images randomly from them. The testing dataset of GC contained additional 50 non-cancer images randomly selected and 50 images of GC picked out before. All the images included were de-identified immediately and then annotated by six endoscopic experts according to a back-to-back protocol. **(B)** A two-stage principle of Faster R-CNN in lesion location. The first stage is to use feature maps of the last convolution layer to generate candidate ROIs. The second stage is to accomplish lesion recognition, position and classification.

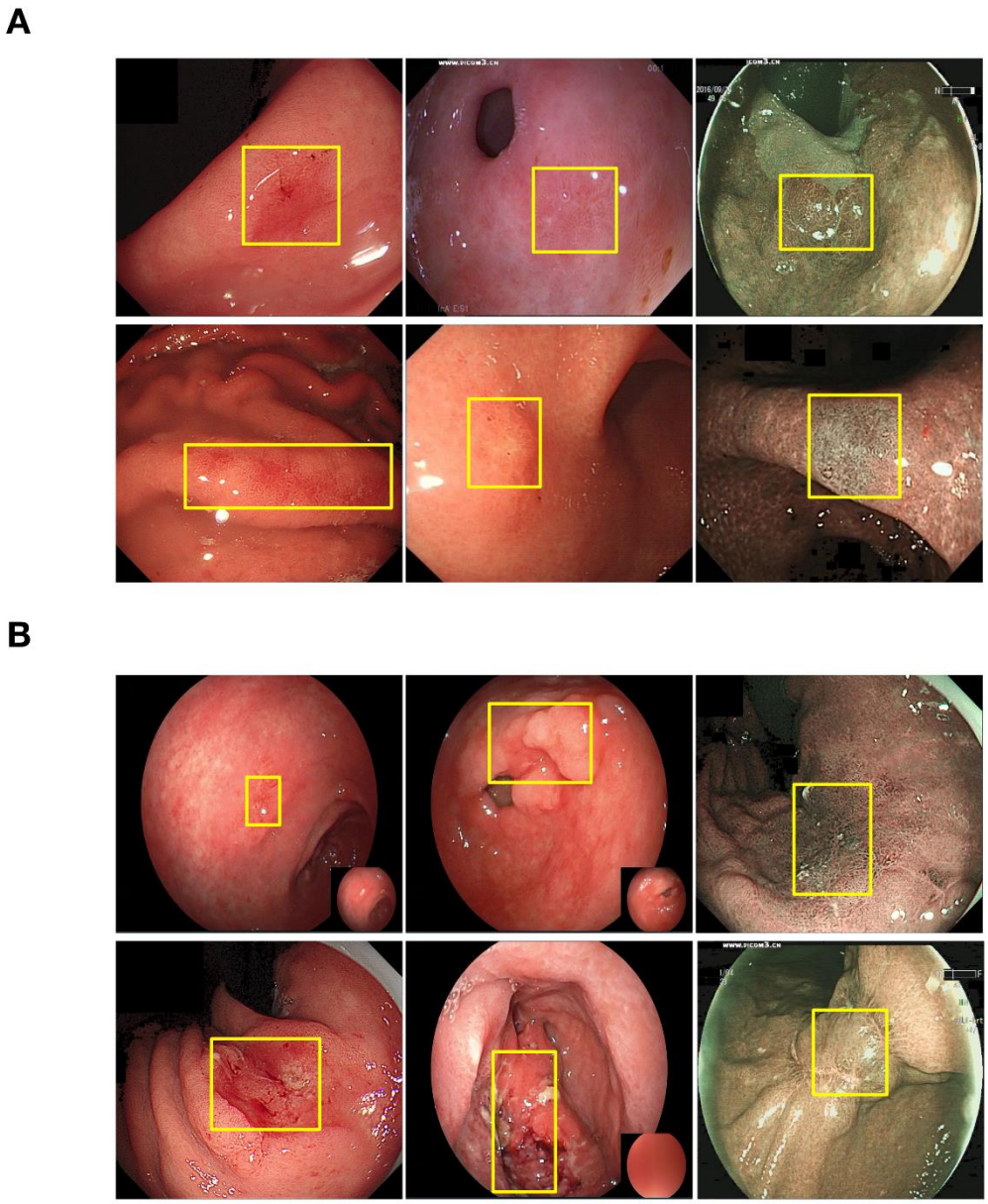

Figure 2. Examples of annotated images for both CAG and GC.

(**A**) Examples of images in the training dataset for CAG with bounding boxes. Each box was located manually in the exact biopsy site according to endoscopic descriptions and histological results, followed by cross-contrast procedure. (**B**) Examples of images in the training dataset for GC with bounding boxes. The images were annotated manually with the boxes in the biopsy sites and as large as possible without exceeding the boundary of the lesions.

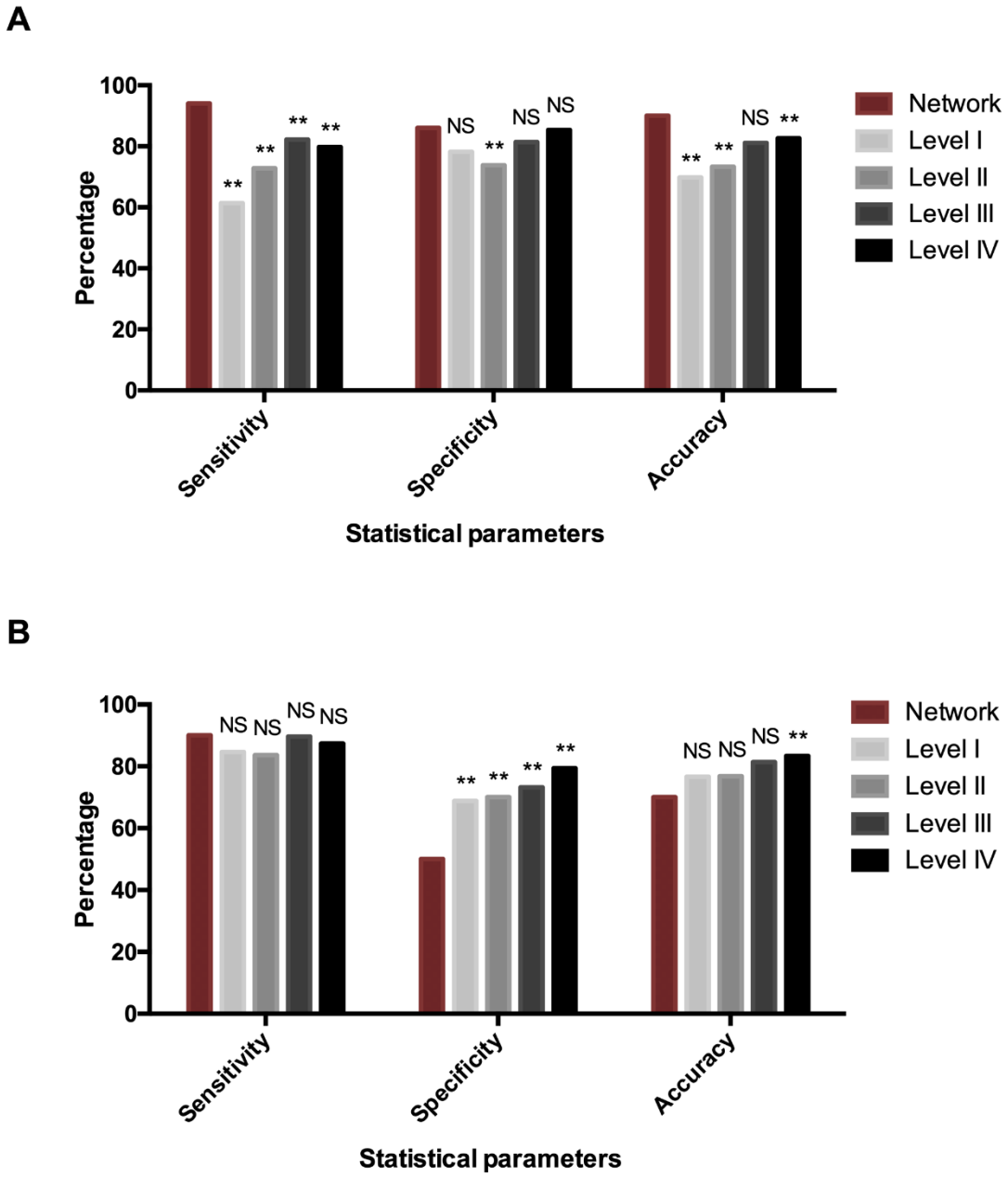

Figure 3. Comparison of performance between network and doctors in different levels in the CAG and GC test.

(**A**) Comparison of diagnostic reliability between the best model and doctors in each level in the CAG test. There is significant difference in sensitivity between the network and doctors from Level I to Level IV ($P < 0.001$, $\chi^2 = 31.226$; $P < 0.001$, $\chi^2 = 16.004$; $P < 0.001$, $\chi^2 = 13.820$; $P = 0.003$, $\chi^2 = 8.665$, respectively). No significant difference in specificity is observed except Level II ($P = 0.141$, $\chi^2 = 2.168$; P = 0.034, $\chi^2 = 4.500$; $P = 0.341$, $\chi^2 = 0.907$; $P = 0.841$, $\chi^2 = 0.040$, respectively). As for accuracy, statistical

difference is observed except Level III ($P < 0.001$, $\chi^2 = 12.500$; $P = 0.002$, $\chi 2 = 9.584$; $P = 0.103$, $\chi^2 = 2.658$; $P = 0.071$, $\chi^2 = 3.267$, respectively). (**B)** Comparison of diagnostic reliability between the best model and doctors in each level in the GC test. No significant difference in sensitivity is observed between the optimal network and doctors from Level I to Level IV ($P = 0.285$, $\chi^2 = 1.143$; $P = 0.207$, $\chi^2 = 1.592$; $P = 1.000$, $\chi^2 = 0.000$; $P = 0.506$, $\chi^2 = 0.442$, respectively). There is statistical difference in specificity comparing the network and doctors in all the four levels ($P = 0.006$, $\chi^2 = 7.490$; $P = 0.004$, $\chi^2 = 8.333$; $P = 0.001$, $\chi^2 = 11.171$; $P < 0.001$, $\chi^2 = 18.364$, respectively). No significant difference in accuracy is observed except Level IV ($P = 0.262$, $\chi^2 = 1.258$; $P = 0.262$, $\chi^2 = 1.258$; $P = 0.071$, $\chi^2 = 3.271$; $P = 0.030$, $\chi^2 = 4.700$, respectively).

**, statistical difference; NS, no statistical difference.