Simultaneous Recognition of Atrophic Gastritis and Intestinal Metaplasia on White Light Endoscopic Images Based on Convolutional Neural Networks: A Multicenter Study

Ne Lin, BM[1], Tao Yu, BE[2], Wenfang Zheng, BM[1,3,4], Huiyi Hu, BE[2], Lijuan Xiang, MM[5], Guoliang Ye, MD[6], Xingwei Zhong, BM[7], Bin Ye, MD[8], Rong Wang, MD[9], Wanyin Deng, MM[10], JingJing Li, MM[11], Xiaoyue Wang, PhD[12], Feng Han, BM[13], Kun Zhuang, MM[14], Dekui Zhang, MM[15], Huanhai Xu, MM[16], Jin Ding, MM[17], Xu Zhang, BE[2], Yuqin Shen, BM[1], Hai Lin, MM[18], Zhe Zhang, MM[19], John J. Kim, MD, MS[20], Jiquan Liu, PhD[2], Weiling Hu, PhD[1,3,4], Huilong Duan, PhD[2] and Jianmin Si, MD[1,3,4]

INTRODUCTION: **Patients with atrophic gastritis (AG) or gastric intestinal metaplasia (GIM) have elevated risk of gastric adenocarcinoma. Endoscopic screening and surveillance have been implemented in high incidence countries. The study aimed to evaluate the accuracy of a deep convolutional neural network (CNN) for simultaneous recognition of AG and GIM.**

METHODS: **Archived endoscopic white light images with corresponding gastric biopsies were collected from 14 hospitals located in different regions of China. Corresponding images by anatomic sites containing AG, GIM, and chronic non-AG were categorized using pathology reports. The participants were randomly assigned (8:1:1) to the training cohort for developing the CNN model (TResNet), the validation cohort for fine-tuning, and the test cohort for evaluating the diagnostic accuracy. The area under the curve (AUC), sensitivity, specificity, and accuracy with 95% confidence interval (CI) were calculated.**

RESULTS: **A total of 7,037 endoscopic images from 2,741 participants were used to develop the CNN for recognition of AG and/or GIM. The AUC for recognizing AG was 0.98 (95% CI 0.97–0.99) with sensitivity, specificity, and accuracy of 96.2% (95% CI 94.2%–97.6%), 96.4% (95% CI 94.8%–97.9%), and 96.4% (95% CI 94.4%–97.8%), respectively. The AUC for recognizing GIM was 0.99 (95% CI 0.98–1.00) with sensitivity, specificity, and accuracy of 97.9% (95% CI 96.2%–98.9%), 97.5% (95% CI 95.8%–98.6%), and 97.6% (95% CI 95.8%–98.6%), respectively.**

DISCUSSION: **CNN using endoscopic white light images achieved high diagnostic accuracy in recognizing AG and GIM.**

SUPPLEMENTARY MATERIAL accompanies this paper at http://links.lww.com/CTG/A653

Clinical and Translational Gastroenterology 2021;12:e00385. https://doi.org/10.14309/ctg.0000000000000385

[1]Department of Gastroenterology, Sir Run Run Shaw Hospital, Medical School, Zhejiang University, Hangzhou, China; [2]Key Laboratory for Biomedical Engineering of Ministry of Education, College of Biomedical Engineering & Instrument Science, Zhejiang University, Hangzhou, China; [3]Zhejiang University Cancer Center, Hangzhou, China; [4]Institute of Gastroenterology, Zhejiang University (IGZJU), Hangzhou, China; [5]Department of Gastroenterology, The People's Hospital of Shangyu, and Shangyu Hospital of the Second Affiliated Hospital, Medical School, Zhejiang University, Shaoxing, China; [6]Department of Gastroenterology, The Affiliated Hospital, Medical School, Ningbo University, Ningbo, China; [7]Department of Gastroenterology, The People's Hospital of Deqing, Huzhou, China; [8]Department of Gastroenterology, The Central Hospital of Lishui City, Lishui, China; [9]Digestive Endoscopy Center, Shanxi Provincial People's Hospital, Taiyuan, China; [10]Department of Gastroenterology, Fujian Provincial Hospital, Fuzhou, China; [11]Department of Gastroenterology, The First People's Hospital of Huzhou, Huzhou, China; [12]Department of Gastroenterology, Beijing Anzhen Hospital, Capital Medical University, Beijing, China; [13]Department of Gastroenterology, The First People's Hospital of Jiaxing, Jiaxing, China; [14]Department of Gastroenterology, The Central Hospital of Xi'an, Xi'an, China; [15]Department of Gastroenterology, The Second Affiliated Hospital, Medical School, Lanzhou University, Lanzhou, China; [16]Department of Gastroenterology, The People's Hospital of Yueqing, Wenzhou, China; [17]Department of Gastroenterology, The Central Hospital of Jinhua, Jinhua, China; [18]Department of Gastroenterology, The People's Hospital of Quzhou, Quzhou, China; [19]Department of Gastroenterology, The People's Hospital of Longyou, Quzhou, China; [20]Division of Gastroenterology, Loma Linda University Health, Loma Linda, California, USA. **Correspondence:** Weiling Hu, MD. E-mail: huweiling@zju.edu.cn. Jiquan Liu, PhD. E-mail: liujq@zju.edu.cn.

Received December 5, 2020; accepted June 16, 2021; published online August 3, 2021

INTRODUCTION

Chronic inflammation of gastric mucosa can evolve to loss of glands, intestinal metaplasia, dysplasia, and eventually gastric adenocarcinoma over several years to decades (1). With goals of identifying patients with precancerous lesions or early detection of gastric cancer, endoscopic screening for general population has been implemented in high incidence countries (2). Furthermore, periodic endoscopic surveillance is generally advised for patients with atrophic gastritis (AG) and gastric intestinal metaplasia (GIM). However, given the poor correlation between endoscopic impression and histological diagnosis of AG and GIM, topographic biopsies are recommended for diagnosis and surveillance of premalignant gastric lesions, especially in China (3–5).

Over the past decade, artificial intelligence technology has made great progress in gastrointestinal endoscopy. Convolutional neural network (CNN), architecture of deep learning for image features analysis, has been applied to improve the diagnosis of various gastrointestinal lesions, such as colorectal polyps, esophageal cancer, *Helicobacter pylori* infection, and gastric cancer (6). Specific endoscopic features from a large data set can be extracted by using multiple network layers and a backpropagation algorithm to develop a model to provide a probability for presence of pathology. Given the remarkable visual recognition capability, we hypothesized that a CNN technology can accurately recognize AG and GIM and guide physicians with biopsies during conventional endoscopy. We have developed a Computer-aided Decision Support System that use CNN for recognition of AG and GIM based on white light endoscopic images. The aim of the study was to evaluate the accuracy of CNN for simultaneous recognition of AG and GIM based on archived endoscopic images.

METHODS

Patients

Patients who received outpatient endoscopy with gastric biopsies between January 2015 and October 2015 with or without upper GI symptoms were evaluated from 14 centers. Gastric biopsies served as the gold standard for AG and GIM. Exclusion criteria included patients with a history of gastric cancer, dysplasia, submucosal tumor, and gastric surgery or incomplete records. The patients meeting inclusion criteria were first divided into pathologic lesions (AG or GIM) or nonpathologic lesions (chronic non-AG). After excluding images with suboptimal quality (i.e., blurred, excessive mucous, light reflections, and excessive bubbles), each image was carefully annotated by anatomic location by reviewing clinical diagnosis, endoscopic findings, and pathology report. Meanwhile, inconsistent images and cases were further excluded manually.

Examinations were performed by experienced endoscopists with biopsies generally obtained from areas of focal mucosal changes or areas suspicious for AG, IM, or dysplasia per practice typical in an area with high incidence of gastric cancer. In the absence of suspicious lesions, routine biopsies of the antrum, corpus, and fundus were obtained per the discretion of the endoscopist. Pathologically confirmation was performed in all gastric biopsy specimens obtained to evaluate for AG and GIM based on the Sydney System Classification of Chronic Gastritis. Atrophy of the gastric mucosa is defined as loss of glandular tissue. Metaplastic epithelium can be recognized morphologically by the presence of goblet cells, absorptive cells, and cells resembling colonocytes or by its enzyme or mucin content (7). The endoscopy images of the study population randomly assigned (8:1:1) to the training, validation, and testing group. The study was approved by the Ethics Committee of the participating hospitals (20190122-8) before initiating the study. Endoscopic images from patients were stored in the retrospective databases at each participating hospital, and informed consent was exempted by the institutional review boards.

Data

The archived endoscopic images obtained using standard white light endoscope (EVIS GIF-Q260, GIF-H260, GIF-H260Z, GIF-H260J, GIF-H290, and GIF-HQ290; Olympus, Tokyo, Japan) from the endoscopic databases were extracted. All endoscopic images were manually labeled by 3 experts. For the annotation of the images, we developed a Web-based platform that contained each patient's clinical diagnosis, procedure report, endoscopy images, and pathology report uploaded by the participating centers. Each image was carefully annotated based on the available information by an expert endoscopist. For anatomic sites containing histologic AG or GIM, boundaries of the focal lesions outlined by the expert endoscopist were used to facilitate accurate annotation of images. Selected images were randomly mirror flipped for data augmentation to improve the accuracy of the model trained by CNN.

CNNs

The CNN model utilized TResNet (Alibaba DAMO Academy, Hangzhou, China) (8) to develop training strategies. Given potential coexistence of AG and GIM, multilabel classification where an instance may be associated with multiple labels was adopted. Fed with 1 endoscopic image of upper gastrointestinal tract, the CNN model outputs a continuous number between 0 and 1 for the probability of chronic non-AG, AG, and GIM, respectively (Figure 1). A validation cohort was used to further fine-tune the CNN model and determine the optimal cutoff values. After deep learning model was constructed, an independent testing cohort was used to evaluate the diagnostic accuracy of CNN for recognition of AG/GIM. Finally, 5-fold stratified cross-validation was used to validate the robustness of the optimized model. For each round, data sets were split into training, validation, and testing sets (80%, 10%, and 10%, respectively). The architecture of CNN model and workflow of the experiment are shown in Figure 2.

Visualization of the CNN models

To validate the interpretability of our optimized model, global average pooling was used to generate class activation maps, which illustrate the weight distribution of feature maps on determination of specific lesion class (9). The output of global average pooling represents the weight coefficient of the image regions, and value of the output can be projected to the feature maps that were extracted from the raw image by CNN. Every single pixel on the feature maps can be traced back to the specific receptive field of input image. Therefore, the raw image was colored along with ratio of neuron weights to display the most suspicious area on the feature map activated by the lesion class.

Statistics

Demographic data were expressed as mean with SD or median with ranges as appropriate. Optimal cutoff values to obtain the highest performance were calculated using Youden index in the validation cohort. Sensitivity, specificity, and accuracy with 95% confidence

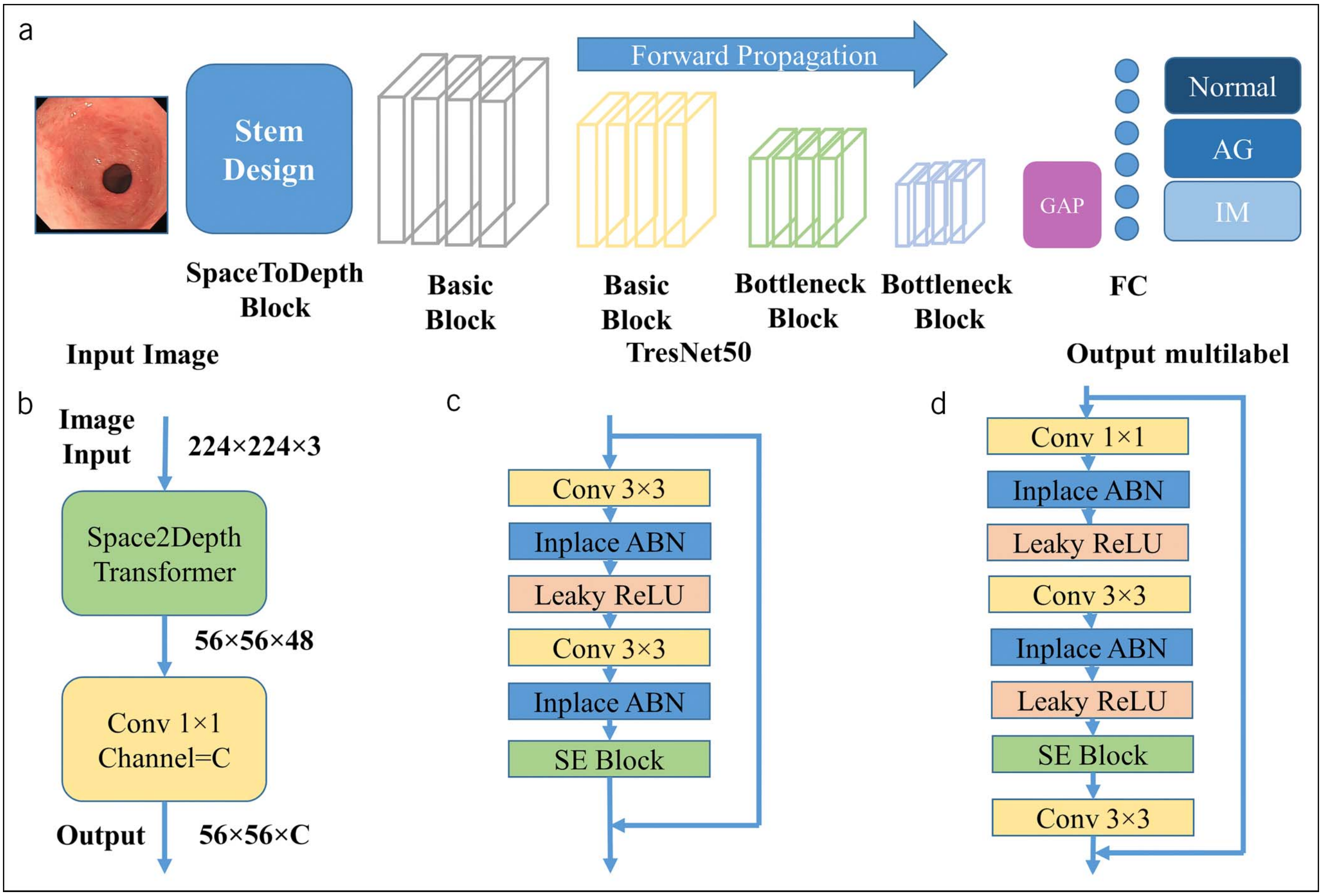

Figure 1. (**a**) The architecture of TResNet used for training the multilabel gastric lesion classification model. It is composed of Stem Design, blocks of residual structure, GAP with a FC. Normal represents gastric mucosa without lesions of atrophy and intestinal metaplasia. (**b**) The structure of stem Design of model entrance in (**a**). It consists of SpaceToDepth transformer block to reduce the loss of information and 1 × 1 convolutional layers to adjust channels. (**c**) The fundamental building structure of Basic Block in (**a**). It consists of convolutional layers, Inplace Activated Batch Normalization operations, Nonlinear Leaky ReLU activation functions, SE attention Block, and skip connections. (**d**) The Bottleneck Block of the experiment in the study, which had similar structure with Basic Blocks but equipped with 1 × 1 Convolutional layer to enhance GPU. AG, atrophic gastritis; FC, fully connected layer; GAP, Global Average Pooling; GIM, gastric intestinal metaplasia; GPU, graphics processing unit usage.

interval (CI) of the CNN model were calculated by using pathologically diagnosis as the gold standard (10). Receiver operating characteristic curves were plotted (Figure 5), and area under the curve (AUC) with 95% CI was calculated. Furthermore, subgroup analysis evaluating sensitivity, specificity, and accuracy by anatomical locations (antrum, angularis, and corpus/fundus) were calculated. SPSS 24.0 (IBM, Armonk, NY) was used for all statistical analysis; a *P* value of <0.05 was considered statistically significant.

RESULTS

Patient characteristics

A total of 16,511 (87.9%) patients received outpatient endoscopy from the largest center (Sir Run Run Shaw Hospital) in the study. Among those, 262 (1.6%) did not receive gastric biopsies, including 254 (1.5%) who had structural abnormality not meeting study criteria and 8 (0.05%) with elevated risk of bleeding. Finally, 1,826 patients from Sir Run Run Shaw Hospital (66.7%) and 915 (33.3%) patients from other 13 hospitals were enrolled (Figure 3). To evaluate for differences in patients enrolled vs not enrolled in the study, patients from the largest center (Sir Run Run Shaw Hospital, Hangzhou, China) were analyzed. The proportion of sex and age distribution between enrolled and not enrolled from Sir Run Run Shaw Hospital were similar (see Supplementary Data Table S1, http://links.lww.com/CTG/A653).

Clinical characteristics of included patients are summarized in Table 1. A total of 2,741 patients including 7,037 archived images were available for the study. The mean age of 2,741 patients was 52.0 ± 13.2, with 1,387 (50.6%) male patients. There were 3,082 images confirmed pathologically as AG and 3,317 images as GIM; of these, 2,899 images were pathologically confirmed as coexistence of AG and GIM.

To evaluate for possible variability of annotation, 3 expert endoscopists analyzed 529 randomly selected images containing AG and/or GIM. The 3 expert endoscopists demonstrated high degree of agreement of annotation of lesions containing AG (kappa value 0.87–0.94) or GIM (kappa value 0.85–0.98) by anatomic sites (see Supplementary Data Table S2, http://links.lww.com/CTG/A653).

Finally, participants were split into 3 parts with 8:1:1 ratio: 2,193 (80%), including 1,173 (53%) patients with AG/GIM (3,076 images), were assigned to the training cohort; 275 (10%), including 147 (53%) patients with AG/GIM (226 images), were

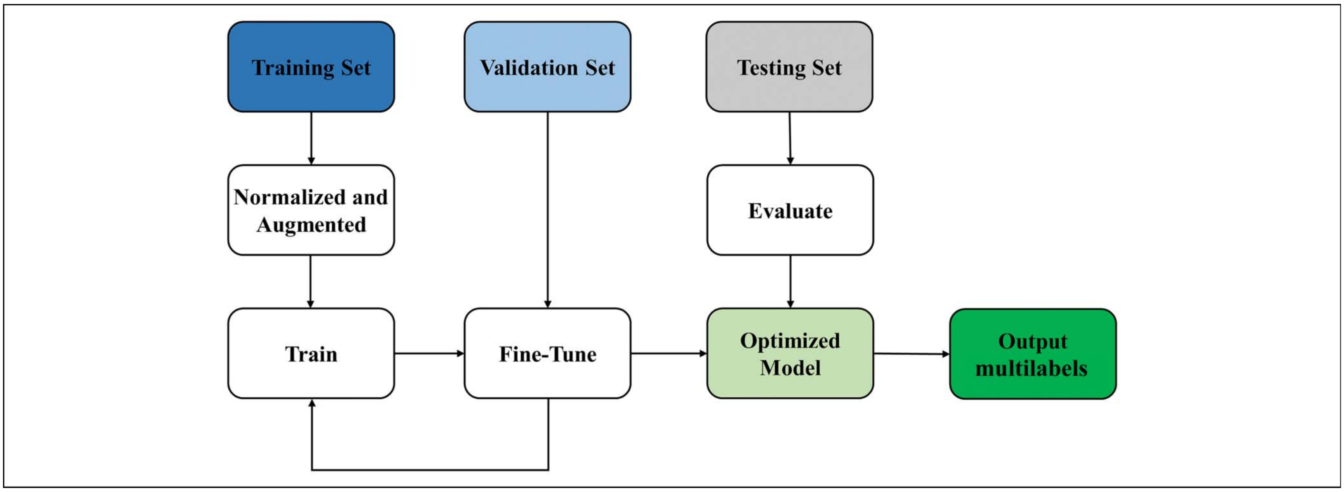

Figure 2. The workflow of the experiment in the study.

assigned to the validation cohort; and 273 (10%), including 147 (54%) patients with AG/GIM (198 images), were assigned to the testing cohort.

Performance of CNN

The diagnostic capability of different fundamental and advanced frameworks of CNN for recognition of AG and/or GIM was evaluated, and TResNet was finally used as representative model (see Supplementary Data Figure S1 and Table S3, http://links.lww.com/CTG/A653). Optimal cutoff values were determined by applying the pretrained CNN model (TResNet) in the validation set. Using an optimal cutoff value of 0.47 in the testing set, the AUC for recognition AG was 0.983 (95% CI 0.967–0.991) with sensitivity, specificity, and accuracy of 96.2% (95% CI 94.2%–97.6%), 96.4% (95% CI 94.8%–97.9%), and 96.4% (95% CI 94.4%–97.8%), respectively. The AUC for recognizing GIM was 0.990 (95% CI 0.976–0.996) with sensitivity, specificity, and accuracy of 97.9% (95% CI 96.2%–98.9%), 97.5% (95% CI 95.8%–98.6%), and 97.6% (95% CI 95.8%–98.6%), respectively, using an optimal cutoff value of 0.16. Cross-validation was performed by splitting the testing set to 5-fold. The average diagnostic accuracy of CNN for recognition of AG and GIM was 96.2% (range 95.8%–96.5%) and 97.2% (range 96.3%–97.6%), respectively (Table 2).

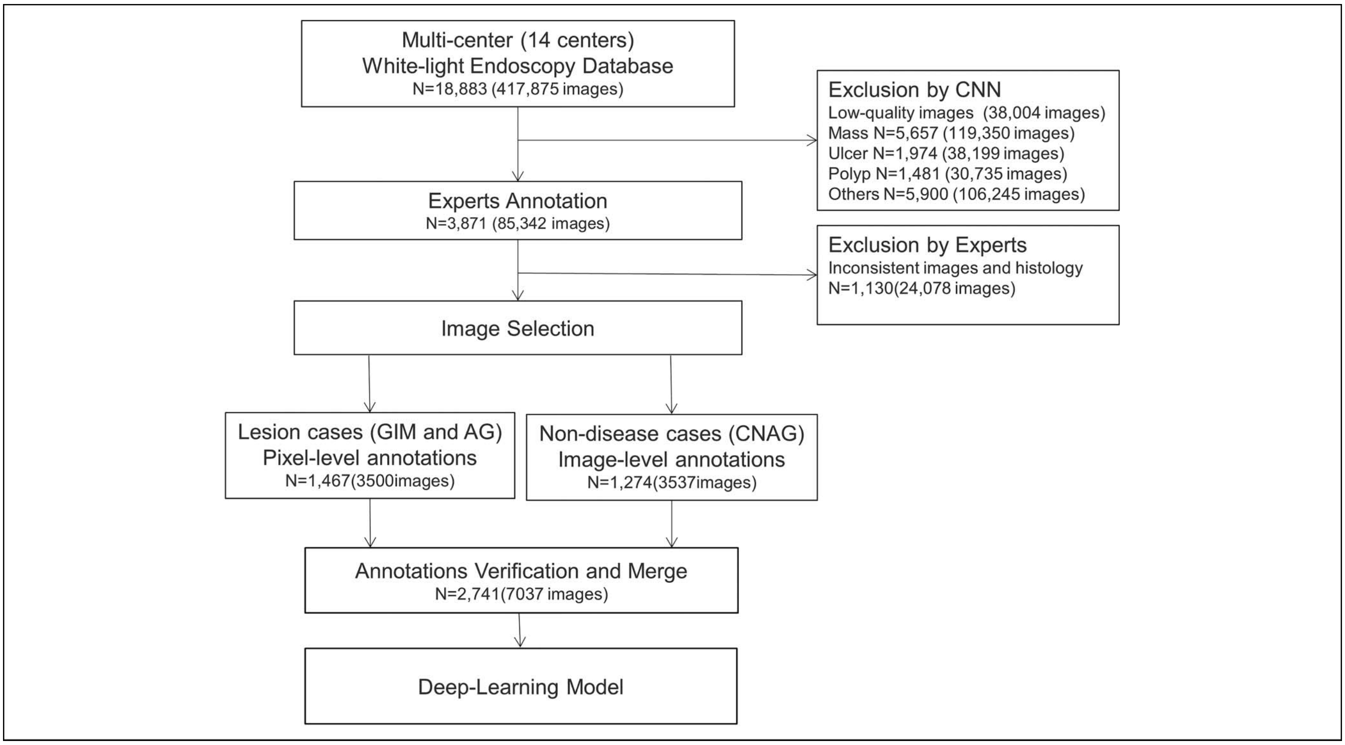
Multi-center (14 centers)
White-light Endoscopy Database
N=18,883 (417,875 images)
Exclusion by CNN
Low-quality images (38,004 images)
Mass N=5,657 (119,350 images)
Ulcer N=1,974 (38,199 images)
Polyp N=1,481 (30,735 images)
Others N=5,900 (106,245 images)
Experts Annotation
N=3,871 (85,342 images)
Exclusion by Experts
Inconsistent images and histology
N=1,130(24,078 images)
Image Selection
Lesion cases (GIM and AG)
Pixel-level annotations
N=1,467(3500images)
Non-disease cases (CNAG)
Image-level annotations
N=1,274(3537images)
Annotations Verification and Merge
N=2,741(7037 images)
Deep-Learning Model

Figure 3. Workflow diagram for data collection, screening, annotation, development and evaluation of AG, GIM and CNAG. AG, atrophic gastritis; CNAG, chronic nonatrophic gastritis; GIM, gastric intestinal metaplasia.

Table 1. Clinical characteristics of included patients

Characteristic	Patients, n			
	All	Training	Validation	Testing
Total	2,741	2,193	275	273
Age, yrs (SD)	52.0 (13.2)	52.2 (13.1)	51.3 (13.9)	51.6 (13.7)
Male patients (%)	1,387 (50.6)	1,120 (51.0)	132 (48.0)	135 (49.3)
	Images, n			
Atrophic gastritis				
Mild	1,177	1,014	87	76
Moderate	1,434	1,287	77	70
Severe	481	412	35	34
Intestinal metaplasia				
Mild	1,002	856	78	68
Moderate	1,483	1,344	79	60
Severe	816	712	44	60
Location				
Fundus	18	17	0	1
Corpus	440	376	36	28
Antrum	4,994	4,164	438	392
Angularis	1,585	1,326	132	127
Original gastric images	7,037	5,883	606	548
Images after data augmentation	54,416	53,262	606	548

Performance of CNN by different anatomic location

When evaluating gastric image by different anatomic location, the AUCs for recognition of AG ranged from 0.974 (95% CI 0.956–0.991) in the antrum to 0.997 (95% CI 0.989–1.00) in the corpus/fundus, whereas the AUCs for recognition of GIM ranged from 0.972 (95% CI 0.954–0.991) in the antrum to 0.996 (95% CI 0.988–1.00) in the corpus/fundus (Table 3).

Performance of endoscopists

Three endoscopists blinded to histology results reviewed the images in the testing group to provide endoscopic diagnosis of AG and GIM. The sensitivity for AG ranged 35.2%–51.7%, specificity 69.6%–80.6%, and accuracy 58.9%–70.6%. The sensitivity for GIM ranged 28.2%–47.3%, specificity 86.4%–96.1%, and accuracy 68.8%–86.4% (see Supplementary Data Table S4, http://links.lww.com/CTG/A653). Local endoscopists provided endoscopic diagnosis of AG with sensitivity of 41.9% (95% CI 33.8%–50.3%), specificity of 96.8% (95% CI 92.1%–99.1%), and accuracy of 67.2% (95% CI 61.3%–72.7%) in the testing group (see Supplementary Data Table S5, http://links.lww.com/CTG/A653).

Table 2. Summary of multilabel classification performance for architecture obtained by 5-fold cross-validation

Cross-validation folds	Atrophic gastritis Accuracy (%)	Intestinal metaplasia Accuracy (%)
1	96.4 (94.4–97.8)	97.6 (95.8–98.6)
2	96.5 (95.0–97.7)	97.4 (96.1–98.4)
3	96.1 (94.4–97.5)	97.1 (95.5–98.2)
4	95.8 (93.8–97.4)	97.4 (95.7–98.6)
5	96.4 (94.2–97.9)	96.3 (93.9–97.9)
Average	96.24	97.16

DISCUSSION

Gastric cancer is a common cancer worldwide associated with high morbidity and mortality. AG and intestinal metaplasia are precancerous lesions and intermediate steps in Correa cascade of gastric carcinogenesis (11). GIM is developed from AG and may further progress to dysplasia. An observational cohort study in Sweden evaluated the incidence of gastric cancer among patients with gastric precancerous lesions within 20 years after gastroscopy. In a low-risk Western population, gastric cancer was estimated to develop in approximately 1 in 256 people with normal mucosa, 1 in 85 with gastritis, 1 in 50 with AG, 1 in 39 with intestinal metaplasia, and 1 in 19 with dysplasia (12). Therefore, AG and GIM have been considered an endoscopic target to identify patients who might benefit from surveillance because of the associated risk of gastric cancer and commonly encountered in clinical practice (13). However, the poor correlation between conventional white light endoscopic and histologic diagnosis of AG/IM have led to challenges in detection of precancerous lesions and reliance on endoscopic practice of routine gastric biopsies. Although image-enhanced endoscopy by magnifying endoscopy or narrowband imaging can improve diagnostic accuracy of AG/IM evaluation, the need for additional equipment and learning-curve limit the generalizability in clinical practice.

Emerging studies have highlighted the remarkable performance of CNN in the medical field including computer-aided diagnosis and medical imaging analysis. In gastrointestinal endoscopy, CNN-based diagnostic system has been evaluated for identification of colon polyps (14–16), screening of Barrett esophagus, early esophageal squamous cell carcinoma (17), and assessment of the invasion depth of gastric cancer (18). Given that, we hypothesized that CNN can facilitate endoscopists to accurately evaluate the presence of premalignant lesions and perform biopsies. Recently, a study used 200 images from patients including 100 images with histology-proven AG to construct a CNN model. This study largely evaluated for AG located in the proximal stomach (gastric corpus and fundus) and achieved a diagnostic accuracy of 93% (19). Given the encouraging preliminary results, we developed a CNN deep learning model to evaluate for presence of AG and GIM in a large patient population encompassing 14 centers.

In our work, a multilabel classification model for simultaneously recognition of AG and GIM using 7,037 endoscopic images including 3,500 images with histology proven for recognition of AG/GIM was developed. The sensitivity, specificity, and accuracy of CNN model for evaluating AG were high at 96.2% (95% CI 94.2%), 96.4% (95% CI 94.8%–97.9%), and 96.4% (95% CI 94.4%–97.8%), respectively. The sensitivity, specificity, and accuracy for evaluating GIM were also high at 97.9% (95% CI 96.2%–98.9%), 97.5% (95% CI 95.8%–98.6%), and 97.6% (95% CI 95.8%–98.6%), respectively. When stratified by anatomic site, AUCs for CNN model recognition of AG ranged from 0.98 (95% CI 0.96–0.99) in the antrum to 0.99 (95% CI 0.99–1.00) in the corpus/fundus, whereas AUCs for CNN model recognition of GIM ranged from 0.99 (95% CI 0.98-1.00) in

Table 3. Diagnostic characteristic of gastric image by different anatomic location

Characteristic	All, N = 548 (95% CI)	Antrum, N = 392 (95% CI)	Angularis, N = 127 (95% CI)	Fundus/corpus, N = 29 (95% CI)
Atrophic gastritis[a]				
Accuracy (%)	96.36 (94.4–97.8)	95.9 (93.5–97.7)	98.4 (94.4–99.8)	93.1 (77.2–99.1)
Sensitivity (%)	96.15 (94.21–97.62)	95.1 (92.2–96.9)	100 (97.1–100)	100 (88.1–100)
Specificity (%)	96.44 (94.8–97.9)	96.4 (94.1–98.0)	97.8 (93.2–99.5)	96.6 (82.2–99.9)
AUC	0.983 (0.967–0.991)	0.981 (0.960–0.991)	0.988 (0.944–0.99)	0.987 (0.989–1.00)
Intestinal metaplasia[b]				
Accuracy (%)	97.6 (95.8–98.6)	97.7 (95.7–98.9)	97.6 (93.2–99.5)	96.6 (82.2–99.9)
Sensitivity (%)	97.9 (96.2–98.9)	98.6 (96.7–99.4)	97.5 (93.2–99.5)	80 (60.3–92.0)
Specificity (%)	97.5 (95.8–98.6)	97.2 (95.0–98.6)	97.7 (94.4–99.8)	100 (88.1–1)
AUC	0.990 (0.976–0.996)	0.990 (0.974–0.997)	0.998 (95.7–1.00)	1.00 (0.881–1.00)

AUC, area under curve; CI, confidence interval.

[a] Optimal accuracy, sensitivity, and specificity using optimal cutoff using 0.47.

[b] Optimal accuracy, sensitivity, and specificity using optimal cutoff using 0.16.

the antrum to 1.00 (95% CI 0.88–1.00) in the corpus/fundus. The high sensitivity of CNN model for ruling out premalignant gastric lesions in our study is especially noteworthy. Consistent with previous studies, our results demonstrated poor sensitivity of endoscopic diagnosis of AG by either expert (41.2%, 95% CI 30.9%–51.5%) or local endoscopists (41.9%, 95% CI 33.8%–50.3%). Similarly, the sensitivity of ruling out endoscopic diagnosis of GIM using static images by expert endoscopist was also poor at 39.2% (95% CI 28.0%–50.4%). However, CNN model demonstrated high diagnostic sensitivities of 96.2% (95% CI 94.2%–97.6%) and 97.9% (95% CI 96.2%–98.9%), for ruling out AG and GIM, respectively.

Recent advancements in artificial intelligence technology have propelled the ability to visualize the internal representation learned by CNN deep learning models to better understand the learned properties. Zeiler and Fergus (20) proposed a multilayered deconvolutional network model to project the feature activations back to the input pixel space for visualizing input stimuli that excite individual feature maps at any layer in the model. Mahendran and Vedaldi (21) conducted a visual analysis of internal representations at different layers by inverting them. Kim et al. constructed a novel method of localizing and visualizing a region of interest within a medical image that is considered to be the most discriminant. The visualization technology provided explanation of the CNN deep learning model predictions to facilitate adoption in clinical settings (22). In our study, optimized class activation mapping was used to localize class-specific image

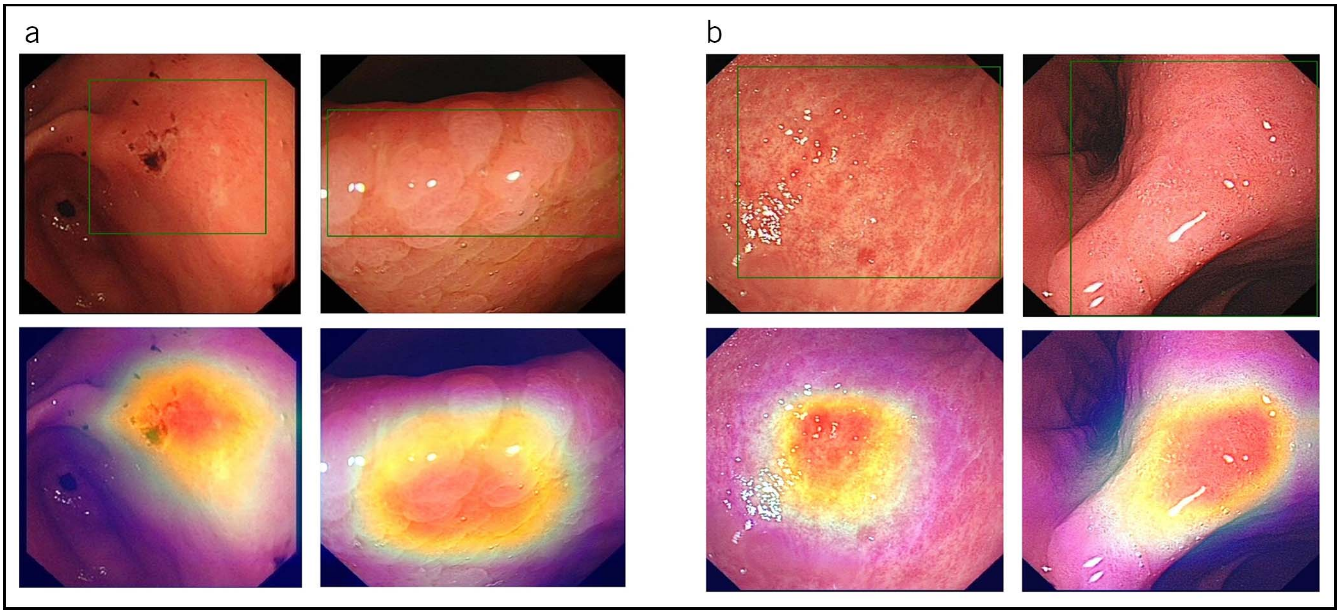

Figure 4. CNN-derived visualization technology localizing of (**a**) AG and (**b**) AG and GIM. The red regions of heat maps are most suspicious area evaluated by the CNN model. AG, atrophic gastritis; CNN, convolutional neural network; GIM, gastric intestinal metaplasia.

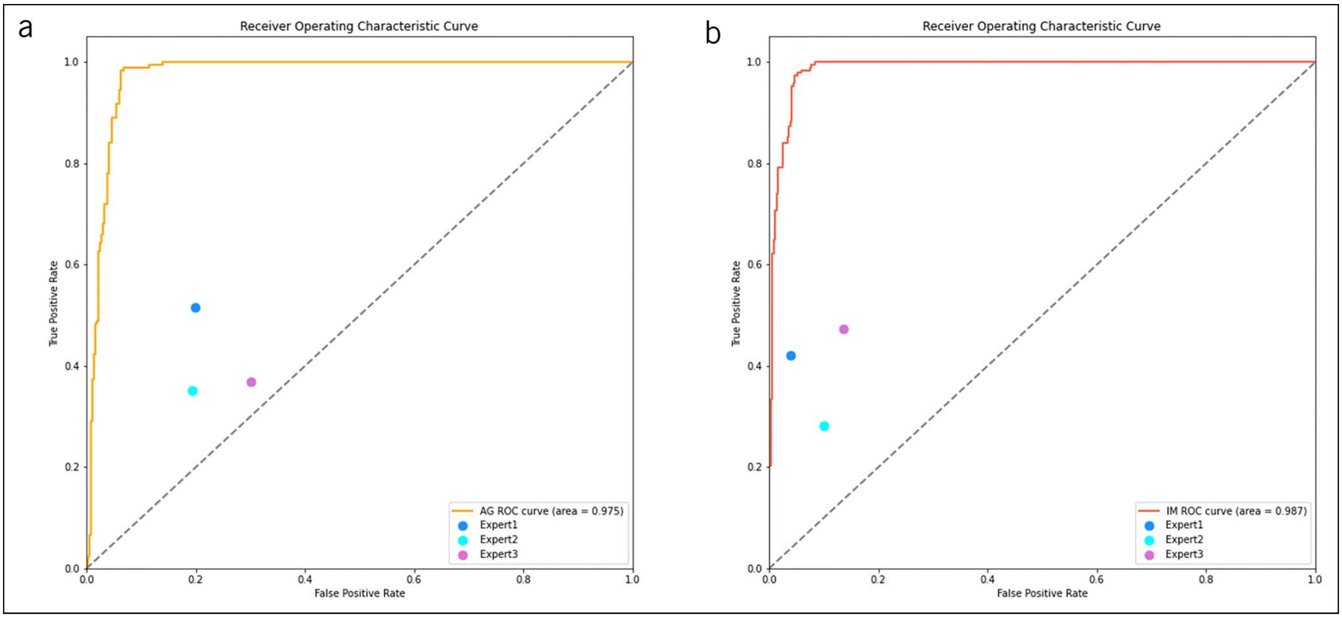

Figure 5. ROC for AG and GIM of optimized CNN model and 3 expert endoscopists, separately. AG, atrophic gastritis; CNN, convolutional neural network; GIM, gastric intestinal metaplasia; ROC, receiver operating curve.

regions in a single forward pass (Figure 4). Red regions of heat maps offered explanation for the classification judgment used by the CNN deep learning model. Furthermore, the impressive correlation between the boundaries of diseased lesion outlined by experts and red regions of heat maps demonstrated the optimized model's capability to accurately recognize and interpret the region of interest. Reverse engineering the predictions derived from machine learning algorithm may lead to better understanding of disease pathogenesis or practice-changing insights. In addition to distinguishing images with or without the presence of AG or GIM, CNN model using visualization technology may provide a region of interest containing premalignant gastric lesion within an image to facilitate targeted biopsies. For example, focal areas of AG or GIM delineated by CNN model using visualization technology correlated with boundaries of AG or GIM annotated by expert endoscopists (Figure 4). The accuracy of CNN model using visualization technology provides direction for future research.

Our findings have clinical implication. Although our study showed that endoscopists are excellent at recognizing the presence of premalignant gastric lesions, the sensitivity to rule out the presence of AG and GIM were poor. The high sensitivity of CNN model for AG and GIM, superior to endoscopic impression even by expert endoscopists, highlights an opportunity to bridge an important gap in clinical practice. If CNN model can accurately rule out the presence of premalignant lesion with high sensitivity, routine biopsies for a large number of patients receiving endoscopy for gastric cancer screening may be obviated. For example, in our study, nearly all (>99%) patients receiving outpatient endoscopy received gastric biopsies, typical of endoscopic practice pattern in area with high incidence of gastric cancer. Furthermore, given high interobserver variability for grading of the severity of gland loss and mucosal changes, the Operative Link on Gastritis Assessment and Operative Link in Gastritis Assessment based on Intestinal Metaplasia systems were developed for risk classification of gastric cancer (23,24). Although requiring validation, our results demonstrating high accuracy of CNN deep learning model for detection of AG and GIM suggested the possibility of risk stratification of gastric cancer without gastric biopsies, leading to reduction of healthcare resources.

There are limitations to our study. First, the imbalanced anatomic location distribution of selected endoscopic images may potentially disrupt the accuracy for recognition of AG/GIM in the fundus or corpus. Validation using larger images of gastric fundus and corpus may enhance the robustness of our results. For optimal assessment, 5 biopsy specimens are recommended to be taken, 2 from the antrum, 2 from the corpus, and 1 from the incisura angularis (6). However, standardized biopsy protocol was not performed, given the retrospective study design with participation of 14 centers. To reduce bias and model overfitting, large number of images verified by expert endoscopist were collected. Furthermore, multimodel cross-validation (see Supplementary Data Figure S1, http://links.lww.com/CTG/A653) was performed to minimize model overfitting during training of the CNN model. Secondly, our algorithm was not able to accurately stage the severity of AG/GIM, which may partly contribute to interobserver variability. Furthermore, study criteria selecting for characteristic pathologic and nonpathologic lesions may have introduced bias leading to high diagnostic accuracy. Our findings from our retrospective study will require validation in prospective studies, ideally using real-time assessment of gastric mucosa. Finally, although the current speed of image classification had achieved 0.017 s/frame, real-time assessment of AG/GIM during living endoscopy rather than static images will be important for application in clinical practice. In future studies, the diagnostic performance of CNN deep learning model using endoscopic video will be explored.

In conclusion, the CNN system based on endoscopic white light images achieved high sensitivity, specificity, and accuracy for recognition of AG and GIM.

CONFLICTS OF INTEREST

Guarantor of the article: Weiling Hu, MD.

Specific author contributors: Ne Lin, BM, Tao Yu, and Wenfang Zheng contributed equally to this work. W.H., J.L., H.D., and J.S.: conception and design. W.H., L.Z., G.Y., X.Z., B.Y., R.W., W.D., J.J.L., Z.W., F.H., K.Z., D.Z., H.X., J.D., H.L., and Z.Z.: provided the endoscopy images. W.H., N.L., and W.Z.: independently screened and labeled the lesions; J.L., T.Y., H.H., and Z.Z.: trained the CNN models. W.Z. and Y.S.: analysis and interpretation of the data. N.L., T.Y., W.Z., and J.J.K.: drafted the article. All authors edited and critically revised the final version of the manuscript.

Financial support: This work was supported by Public Welfare Research Project of Zhejiang Province (LGF18H160012), Medical Health Project of Zhejiang Province (2020RC064), and National Natural Science Foundation of China (31771072 and 81827804).

Potential competing interests: None to report.

Study Highlights

WHAT IS KNOWN

✓ Accurate detection of precancerous lesions is important for risk stratification of gastric cancer.

✓ Correlation between endoscopic impression and histological diagnosis of AG and gastric intestinal metaplasia (GIM) are poor.

WHAT IS NEW HERE

✓ Convolutional neural network (CNN) accurately recognized AG and GIM.

✓ Diagnostic accuracy of CNN exceeded those of local and experts endoscopists.

TRANSLATIONAL IMPACT

✓ CNN may increase the diagnostic yield of AG and GIM during endoscopy.

✓ CNN may reduce the need for gastric biopsies for stratifying risk of gastric cancer.

REFERENCES

1. Correa P. Human gastric carcinogenesis: A multistep and multifactorial process: First American Cancer Society award lecture on cancer epidemiology and prevention. Cancer Res 1992;52:6735–40.
2. Pimentel-Nunes P, Libânio D, Marcos-Pinto R, et al. Management of epithelial precancerous conditions and lesions in the stomach (MAPS II): European Society of Gastrointestinal Endoscopy (ESGE), European Helicobacter and Microbiota Study Group (EHMSG), European Society of Pathology (ESP), and Sociedade Portuguesa de Endoscopia Digestiva (SPED) guideline update 2019. Endoscopy 2019;51:365–88.
3. Lim JH, Kim N, Lee HS, et al. Correlation between endoscopic and histological diagnoses of gastric intestinal metaplasia. Gut Liver 2013;7:41–50.
4. Eshmuratov A, Nah JC, Kim N, et al. The correlation of endoscopic and histological diagnosis of gastric atrophy. Dig Dis Sci 2010;55:1364–75.
5. Du Y, Bai Y, Xie P, et al. Chronic gastritis in China: A national ulti-center survey. BMC Gastroenterol 2014;14:21.
6. Le Berre C, Sandborn WJ, Aridhi S, et al. Application of artificial intelligence to gastroenterology and hepatology. Gastroenterology 2020; 158:76–94.e2.
7. 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. Am J Surg Pathol 1996 1994;20: 1161–81.
8. Ridnik T, Lawen H, Noy A, et al. TResNet: High Performance GPU-Dedicated Architecture. arXiv e-prints, unpublished results; arXiv: 2003.13630.
9. Zhou B, Khosla A, Lapedriza A, et al. Learning Deep Features for Discriminative Localization. arXiv e-prints, unpublished results; arXiv: 1512.04150.
10. Youden WJ. Index for rating diagnostic tests. Cancer 1950;3(1):32–5.
11. Leung WK, Sung JJ. Review article: Intestinal metaplasia and gastric carcinogenesis. Aliment Pharmacol Ther 2002;16:1209–16.
12. Song H, Ekheden IG, Zheng Z, et al. Incidence of gastric cancer among patients with gastric precancerous lesions: Observational cohort study in a low risk Western population. BMJ 2015;351:h3867.
13. Gupta S, Li D, El Serag HB, et al. AGA clinical practice guidelines on management of gastric intestinal metaplasia. Gastroenterology 2020;158: 693–702.
14. Byrne MF, Chapados N, Soudan F, 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.
15. Zhang R, Zheng Y, Poon CCY, et al. Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker. Pattern Recognit 2018;83:209–19.
16. Urban G, Tripathi P, Alkayali T, et al. Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. Gastroenterology 2018;155:1069–78.e8.
17. Guo L, Xiao X, Wu C, et al. Real-time automated diagnosis of precancerous lesions and early esophageal squamous cell carcinoma using a deep learning model (with videos). Gastrointest Endosc 2020;91:41–51.
18. Zhu Y, Wang QC, Xu MD, et al. Application of convolutional neural network in the diagnosis of the invasion depth of gastric cancer based on conventional endoscopy. Gastrointest Endosc 2019;89:806–15.e1.
19. Guimarães P, Keller A, Fehlmann T, et al. Deep-learning based detection of gastric precancerous conditions. Gut 2020;69:4–6.
20. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. Cham, Springer International Publishing, 2014.
21. Mahendran A, Vedaldi A, IEEE. Understanding deep image representations by inverting them. 2015 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2015:5188-5196.
22. Kim I, Rajaraman S, Antani S. Visual interpretation of convolutional neural network predictions in classifying medical image modalities. Diagnostics (Basel) 2019;9:38.
23. Rugge M, Genta RM. Staging and grading of chronic gastritis. Hum Pathol 2005;36:228–33.
24. Capelle LG, de Vries AC, Haringsma J, et al. The staging of gastritis with the OLGA system by using intestinal metaplasia as an accurate alternative for atrophic gastritis. Gastrointest Endosc 2010;71:1150–8.