Article published online: 2021-05-27



# Expert-level classification of gastritis by endoscopy using deep learning: a multicenter diagnostic trial

OPEN ACCESS

Authors

Ganggang Mu*,1,2,3, Yijie Zhu*,1,2,3, Zhanyue Niu*,4, Hongyan Li1,2,3, Lianlian Wu1,2,3, Jing Wang1,2,3, Renquan Luo1,2,3, Xiao Hu5, Yanxia Li1,2,3, Jixiang Zhang1,2,3, Shan Hu5, Chao Li5, Shigang Ding**,4, Honggang Yu**,1,2,3

Institutions

1 Department of Gastroenterology, Renmin Hospital of Wuhan University, Wuhan, China
2 Key Laboratory of Hubei Province for Digestive System Disease, Renmin Hospital of Wuhan University, Wuhan, China
3 Hubei Provincial Clinical Research Center for Digestive Disease Minimally Invasive Incision, Renmin Hospital of Wuhan University, Wuhan, China
4 Peking University Third Hospital, Beijing, China
5 Wuhan EndoAngel Medical Technology Company, Wuhan, China

submitted 17.9.2020
accepted after revision 14.12.2020

Bibliography
Endosc Int Open 2021; 09: E955–E964
DOI 10.1055/a-1372-2789
ISSN 2364-3722



Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

Corresponding authors
Professor Honggang Yu, Department of Gastroenterology, Renmin Hospital of Wuhan University, 99 Zhangzhidong Road, Wuhan 430060, Hubei Province, China
Fax: 027-88042292
yuhonggang@whu.edu.cn

Supplementary material is available under https://doi.org/10.1055/a-1372-2789

## ABSTRACT

**Background and study aims** Endoscopy plays a crucial role in diagnosis of gastritis. Endoscopists have low accuracy in diagnosing atrophic gastritis with white-light endoscopy (WLE). High-risk factors (such as atrophic gastritis [AG]) for carcinogenesis demand early detection. Deep learning (DL)-based gastritis classification with WLE rarely has been reported. We built a system for improving the accuracy of diagnosis of AG with WLE to assist with this common gastritis diagnosis and help lessen endoscopist fatigue.

**Methods** We collected a total of 8141 endoscopic images of common gastritis, other gastritis, and non-gastritis in 4587 cases and built a DL -based system constructed with UNet++ and Resnet-50. A system was developed to sort common gastritis images layer by layer: The first layer included non-gastritis/common gastritis/other gastritis, the second layer contained AG/non-atrophic gastritis, and the third layer included atrophy/intestinal metaplasia and erosion/hemorrhage. The convolutional neural networks were tested with three separate test sets.

**Results** Rates of accuracy for classifying non-atrophic gastritis/AG, atrophy/intestinal metaplasia, and erosion/hemorrhage were 88.78%, 87.40%, and 93.67% in internal test set, 91.23%, 85.81%, and 92.70% in the external test set ,and 95.00%, 92.86%, and 94.74% in the video set, respectively. The hit ratio with the segmentation model was 99.29%. The accuracy for detection of non-gastritis/common gastritis/other gastritis was 93.6%.

**Conclusions** The system had decent specificity and accuracy in classification of gastritis lesions. DL has great potential in WLE gastritis classification for assisting with achieving accurate diagnoses after endoscopic procedures.

* Contributed equally to this work.

** These authors contributed equally to this work.

## Introduction

Gastric cancer is the fifth most commonly diagnosed malignancy and the third leading cause of cancer-related deaths [1]. Gastritis is related to peptic ulcers and gastric cancer. Gastric cancer develops from superficial gastritis, abd atrophic gastritis (AG), and progressions from metaplasia to dysplasia and carcinoma. Gastric atrophy (GA) and intestinal metaplasia (IM) are the most common stages in gastric carcinogenesis [2, 3]. Many gastric adenocarcinomas are associated with a series of pathological changes caused by long-term gastric mucosa inflammation [4]. Studies suggest that identifying gastric lesions may facilitate early detection of precancerous conditions [5, 6]. Timely detection and treatment of gastritis, especially chronic atrophic gastritis (CAG, including GA and IM), can prevent further deterioration.

Esophagogastroduodenoscopy (EGD) is a routine approach to gastritis diagnosis; however, the accuracy of diagnosis with it varies among endoscopists. Not all endoscopists can diagnose precisely on EGD. The accuracy of CAG endoscopic diagnosis with white-light endoscopy (WLE) reached 0.42 to 0.80 compared with biopsy results [7–9]. To improve the quality of gastritis diagnosis, experts have proposed many guidelines and consensus [10–13]. One study showed that CAG diagnosis accuracy in WLE of endoscopists only reached 46.8 % after guideline-based training [14]. As reported, the accuracy of gastritis diagnosis in WLE was not that good. A system for classifying gastritis lesions in real time is needed [15, 16].

Use of deep learning (DL) technology in artificial intelligence (AI) recently has been introduced in the field of medicine. Deep convolutional neural networks (DCNNs) are being used clinically in a dermatologist-level classification system of skin cancer [17]. The development of AI in EGD also is growing rapidly. Achievements have been made in applying DL to gastritis pathology and systems for X-ray detection [18, 19]. In previous studies, AI has been applied to detection of *Helicobacter pylori*-associated gastritis and AG [8, 9, 20]. Gastric cancer risk stratification system also has been developed [21]. Nevertheless, DCNN-assisted classification of endoscopic gastritis rarely has been studied.

Our team developed a novel system named ENDOANGEL, which uses AI to reduce the blind spot rate with EGD, and conducted a clinical trial to verify its effectiveness and safety [22]. The advantage of the system is that is an AI application designed specifically for use in the gastrointestinal tract [22–24]. Based on our previous study, we aimed to develop a novel real-time DCNN-based system for common gastritis lesion classification and location. This system would result in a summary of photodocumentation at the end of an endoscopic examination.

## Materials and methods

### Study design

We retrospectively collected WLE images to for use in a DL-based, gastritis-assisted diagnostic system. The gold standard for the training and test sets was the consensus of three reviewers regarding non-gastritis and histological results for CAG. We designed this classification system to help recognize and locate gastritis lesions. The system determines the type of lesion based on details observed as the endoscope nears it. Three separate test sets were used for validation. Experts and non-experts from two hospitals participated in three tests of the system.

Three experts participated as reviewers, each of whom had at least 3 years of experience in endoscopy and an annual EGD volume of 1000 to 3000 cases at Renmin Hospital of Wuhan University. The filter criterion was established by three reviewers after face-to-face discussions and used to select images that all of the reviewers all agreed after discussion would guarantee the model's accuracy. There is broad consensus about the basic distinction between AG and non-atrophic gastritis [25–27]. The images showing GA and IM were confirmed with histological results and those that did not require pathology were used after the three reviewers came to consensus about them. Images for GA mean images on which only atrophy was present in pathological results and images for IM means that the reviewers annotated the region in images with IM based pathological results.

Another five endoscopists (not including the reviewers) from our hospital, including four non-experts and one expert, and two experts from Peking University Third Hospital (PUTH) participated in an independent test against the machine. Experts were defined as endoscopists with more than 3 years of experience with EGD and non-experts were defined as endoscopists with less than 1 year of experience with EGD.

### Structure of the system

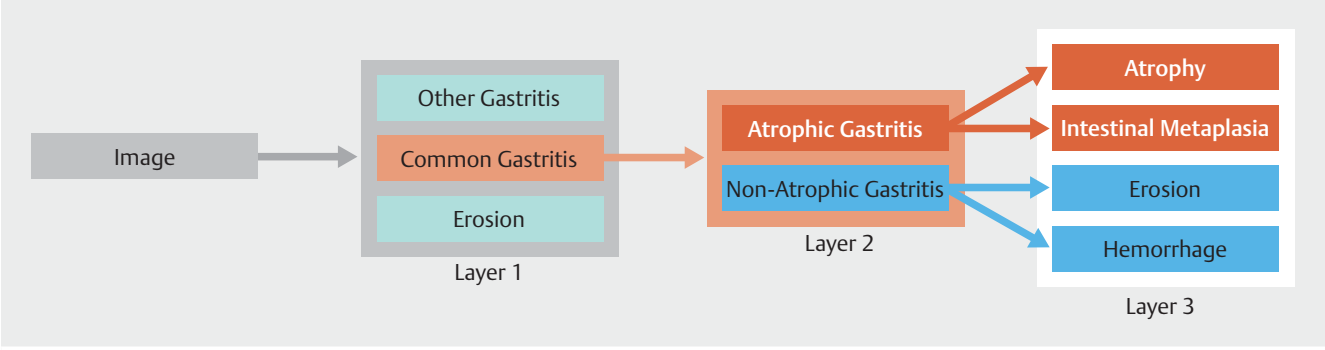
The gastritis lesion classification system was designed to assist the system for diagnosing gastritis. The real-time classification system predicts GA, IM, and erosive and hemorrhagic gastritis, and they are rendered layer by layer. (▶ **Fig. 1**) Images were first classified into non-gastritis, common gastritis, and other gastritis. Non-gastritis meant absence of gastritis. Common gastritis referred tos three kinds of common and meaningful gastritis (classified into AG and non-atrophic gastritis): AG, erosive gastritis, and hemorrhagic gastritis. Other gastritis included bile reflux gastritis and hypertrophic gastritis, cases of which are seen in the authors' hospitals. Images for AG included those with GA and/or IM, while those for non-atrophic gastritis were gastritis images without AG (mainly divided into erosive and hemorrhagic gastritis).

### Preparation of datasets for training and testing

For the training, we collected 8141 WLE images from 4587 patients who had undergone gastroscopy in our hospital between November 2017 and October 2019.

First, we trained and tested our first model (DCNN1) to decide whether the input image was common gastritis. A total of 7326 images were used for training and 815 images were used for validation (**Supplementary Table 2**).

Second, from the data set mentioned above, 5651 common gastritis images were used to train our segmentation model (FCN1) and 1775 non-gastritis images were used as negative samples. Another 570 images selected at random from internal




▶ **Fig. 1** Structure of our system.

and external test sets were used for validation (**Supplementary Table 3**).

Four data sets then were used to train and test our classification models (DCNN2, DCNN3, and DCNN4) in discriminating between types of common gastritis. DCNN2 distinguishes AG from non-AG. DCNN3 classifies GA and IM and DCNN4 includes erosion and hemorrhage. Three separate test sets were prepared for testing, two of which were from our hospital; the other one was from PUTH. A total of 453 images from 386 patients from November 2019 to December 2019 were collected as an internal test set. Furthermore, we collected 80 video clips of 80 cases of four kinds of gastritis in January 2020 as a video set. These original data were captured from standard EGD (CVL-290SL, Olympus Optical Co. Ltd., Tokyo, Japan; VP-4450HD, Fujifilm Co., Kanagawa, Japan). A total of 258 images taken from 137 patients in January 2020 were collected in PUTH as an external test set. The procedure was performed by standard EGD (EG-590WR, EC-590WM, EC-L590ZW, EG-L590ZW, EG-600WR, EC-600WI, EG-601WR and EC-601WI, Fujifilm Co., Kanagawa, Japan). All the images were WLE images. Distribution of the images is shown in ▶ **Table 1**.

## Image preprocessing

Three reviewers came to consensus on criteria about the images. Images that are blurry, dark, out of focus or had mucus and froth were excluded. Two medical doctoral candidates from our hospital trained and supervised by one expert filtered the unqualified images. All personal information was cropped out of the original images.

## Annotation of training set

To ensure that the machine learned the precise characteristics of lesions, single-lesion images were extracted. Three reviewers annotated the training set via an annotation tool (http://www.robots.ox.ac.uk/~vgg/software/via/via-2.0.2.html, VGG Image Annotator (VIA) Abhishek Dutta, Ankush Gupta and Andrew Zisserman). They annotated the dataset together, had a discussion about the controversial images, and then reached a consensus. The resulting classification was added to every extracted lesion.

## Image classification

The DCNN-based system was constructed based on the clinical significance of lesions. Two types of common gastritis – erosive and hemorrhagic – and two premalignant lesions – GA and IM – included (▶ **Fig. 1**).

## Demonstration in videos

To test this model in real clinical practice, 80 video clips from 80 cases of four kinds of gastritis were collected as a video set. The video clips including gastritis lesions (including scope-forward, observing, and scope-withdraw video clips) were clipped by the

▶ **Table 1** Distribution of training and validation set of DCNN2, DCNN3 and DCNN4 (Layer 3).

| | | Hemorrhage | Erosion | Atrophy | IM | Total |
|---|---|---|---|---|---|---|
| Training set | No. images | 880 | 1728 | 1975 | 1068 | 5651 |
| | No. lesions | 968 | 1901 | 2172 | 1175 | 6216 |
| Internal test set | No. images | 80 | 135 | 140 | 98 | 453 |
| | No. lesions | 88 | 149 | 154 | 108 | 499 |
| External test set | No. images | 59 | 65 | 67 | 67 | 258 |
| | No. lesions | 65 | 72 | 74 | 74 | 285 |
| Video set | No. cases | 16 | 22 | 23 | 19 | 80 |

IM, Intestinal Metaplasia.

▶ **Table 2** Baseline information for test sets.

| | Internal test set | External test set | Video set |
|---|---|---|---|
| No. images | 453 | 258 | – |
| No. patients | 386 | 137 | 80 |
| Mean age, y (SD) | 51.74 (11.48) | 53.54 (13.57) | 49.91 (12.93) |
| Sex n (%) | | | |
| ▪ Male | 199 (48.45) | 68 (49.64) | 46 (57.50) |
| ▪ Female | 187 (51.55) | 69 (50.36) | 34 (42.50) |
| Duration (SD) | – | – | 50.95 (31.58) |
| Case classification n (%) | | | |
| ▪ Atrophy | 116 (30.05) | 36 (26.28) | 23 (28.75) |
| ▪ Intestinal metaplasia | 74 (19.17) | 35 (25.55) | 19 (23.75) |
| ▪ Erosion | 120 (31.09) | 33 (24.09) | 22 (27.50) |
| ▪ Hemorrhage | 76 (19.69) | 33 (24.09) | 16 (20.00) |

SD, standard deviation.

three reviewers mentioned above. The average duration of the 80 video clips was 50.95 ± 31.58 seconds. The videos were clipped into images at three frames per second in cases to test the model's stability. The performance of CNNs in videos was evaluated based on the lesions, and a lesion was regarded as correctly predicted when 70 % of the frames were labelled with the correct answer. Similarly, the seven endoscopists from two hospitals completed the answer sheets for the test independently. Screenshots of our real-time gastritis lesion classification system are shown in **Supplementary Fig. 1**.

## Development of the algorithm

We used two kinds of models to construct the gastritis lesion classification system: Unet++ for segmentation and Resnet-50 for classification. Unet++ is a powerful architecture for medical image segmentation and Resnet-50 is a residual learning framework with better ability for generalization [28, 29]. We used transfer learning to train our models [30]. We retrained them using our datasets and fine-tuned the parameters to fit our needs (**Supplementary Table 4**). Dropout, data augmentation, and early stopping were used to decrease the risk of overfitting. The architecture of the CNN is shown in **Supplementary methods and materials** and **Supplementary Fig. 2**.

## Validation of the algorithm

The baseline information for the test sets is shown in ▶ **Table 2**. Images from the same patient did not appear in both the training and test sets. The results from the three reviewers and the histological results were the gold standard in the training and test sets. Another seven endoscopists participated in independent testing, which was compared with results from the system.

## Outcome measurements

### Accuracy of the DCNN-based gastritis classification system

The performances of DCNN1, DCNN2, DCNN3, and DCNN4 are shown separately. Because our study mainly targeted four kinds of common gastritis, the performance of DCNN1 is only reflected in terms of its accuracy. The comparison metrics were accuracy, sensitivity, specificity, positive predicted value (PPV), and negative predicted value (NPV) (**Supplementary methods and materials**).

### Assessment of Unet++

The primary goal of our system was to detect gastritis lesions and sort them clinically rather than precisely describing the exact lesion border. Therefore, pixel-precise delineation metrics were less important for our study and we assessed the accuracy of Unet++ by its hit ratio.

Three reviewers met to assess the hit ratio of the model. They came to consensus on whether the images, after segmentation, contains representative characteristics for classification.

Hit ratio = the number of the representative images/ the total number of dataset.

The Hit ratio was calculated for the unit as an entire image or a single lesion separately. Our results revealed the hit ratio for each kind of gastritis and a total hit ratio.

## Assessment of the location of gastritis lesions

We have developed a DCNN-based system that has been proven to perform better than endoscopists in monitoring blind spots in clinical practice. The location-predict architecture has been proven in a clinical trial and its accuracy is reliably high: 90.02 % in images and 97.20 % in videos [22–24]. The model has matured sufficiently that it can tell the exact location in stomach, and it can identify the specific location of the gastri-

VIDEO

▶ **Video 1** The DCNN-based system show good performance in real EGD videos. Demonstrated videos are atrophy, intestinal metaplasia, erosion and hemorrhage. Live videos are on the left of the screen, while the left is gastritis lesion prediction. Above is real-time classification and below are location-prediction and thumbnail of real-time images with confidence on it. A summary of gastritis lesions will be shown when procedure finished.

tis lesions plus their architecture. A typical video classification system with location prediction is shown in ▶**Video 1**.

### Ethics

This study was approved by the Ethics Committee of Renmin Hospital of Wuhan University and Peking University Third Hospital. Because this was a retrospective study, the Ethics Committees deemed it exempt from a need for informed consent.

### Statistical analysis

We used a two-tailed unpaired Student's *t*-test with a significance level of 0.05 to compare differences in accuracy, sensitivity, specificity, PPV, and NPV of the CNNs and experts. Interobserver agreements of the endoscopists were evaluated using Cohen's kappa coefficient. All analyses were performed with SPSS 26 (IBM, Chicago, Illinois, United States).

## Results

Representative images of four kinds of gastritis lesions are shown in ▶**Fig. 2**. A flowchart for development and evaluation of the system is shown in ▶**Fig. 3**.

Five models were constructed to separately predict the classification of gastritis lesions. The accuracy of DCNN1 was 93.6%, and the separate accuracies for common gastritis, non-gastritis and other gastritis were 95.8%, 88.2%, and 90.3%, respectively. The hit ratio for the segmentation model was 90.96% calculated in lesions and 99.29% in entire images (**Supplementary Table 1**). The performances of DCNN2, DCNN3, DCNN4, and the endoscopists, experts and non-experts are shown in **Supplementary Tables 5, 6,** and **7**. The interobserver agreement for endoscopists is shown in **Supplementary Table 8**.

### Performance of DCNNs and endoscopists in internal test set

In the internal test set, the rates of accuracy of DCNN2, DCNN3, and DCNN4 were 88.78%, 87.40%, and 93.67%, respectively, which is superior to the endoscopists' average level. The accuracy of DCNN2 was significantly higher than that of the seven endoscopists (82.63±6.07%, $P=0.047$). DCNN2 possessed higher sensitivity (88.93%) for identification of AG than did the endoscopists (77.14±10.13%, $P=0.029$). The specificity (88.61% and 88.74±10.10%, $P=0.975$) of recognition of AG between machine and endoscopists was comparable. DCNN3 did better in detecting GA and IM than did the endoscopists (66.89±10.03%, $P=0.02$). The sensitivity and specificity for GA of the machine were 91.56% and 81.48%, respectively, which is higher than for the endoscopists (70.77±11.15%, $P=0.04$ and 61.26±21.55%, $P=0.061$). The accuracy of DCNN4 was higher than that of the endoscopists (82.41±10.79%, $P=0.043$). The accuracy of DCNN2, DCNN3, and DCNN4 was superior to that for the non-experts.

### Performance of DCNNs and endoscopists in external test set

In the external test set, the rates of accuracy of DCNN2, DCNN3, and DCNN4 were 91.23%, 85.81%, and 92.70%, respectively. All were significantly higher than those for the endoscopists (83.54±4.57%, $P=0.006$, 70.91±6.49%, $P=0.001$ and 84.58±5.86%, $P=0.013$). Also, they were higher than for the non-experts (79.99±2.15%, $P=0.003$, 67.03±5.74%, $P=0.011$ and 80.62±3.22%, $P=0.007$). The sensitivity for AG of DCNN2 (93.24%) reached a higher level than did the sensitivity for AG of the endoscopists (78.88±5.66%, $P=0.001$). The sensitivity of DCNN3 (82.43%) in recognizing IM was higher than that for the endoscopists (61.10±15.68%, $P=0.016$). For DCNN4, the machine's sensitivity for erosion reached 95.83%, superior to that for the endoscopists (82.64±9.11%, $P=0.013$). The accuracy of DCNN3 (85.81%) was even higher than that of the experts (79.06±2.71%, $P=0.037$).

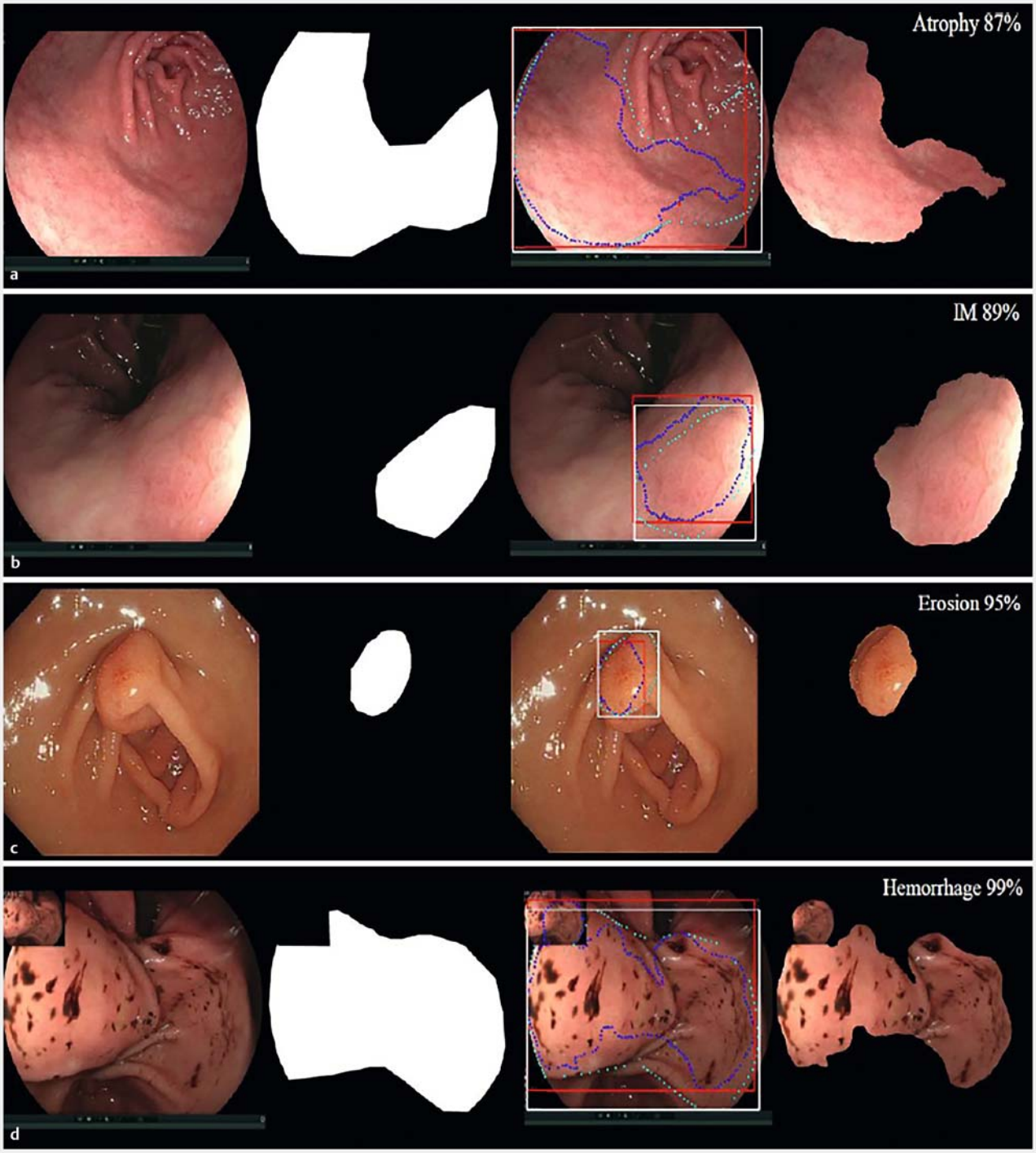
### Performance of DCNNs and endoscopists in real-time videos

In the video set, the accuracies of DCNN2 (95.00%), DCNN3 (92.86%), and DCNN4 (94.74%) were at the same levels as those for the endoscopists (88.21±9.70%, $P=0.138$, 86.05±12.37%, $P=0.227$ and 83.46±12.79%, $P=0.074$). There was no significant difference between CNNs and the endoscopists.

Moreover, we found that the CNNs' capability for recognizing gastritis lesions was comparable to that of experts, as there was no significant difference between experts and CNNs in most cases. A comparison of results is shown in ▶**Fig. 4**.

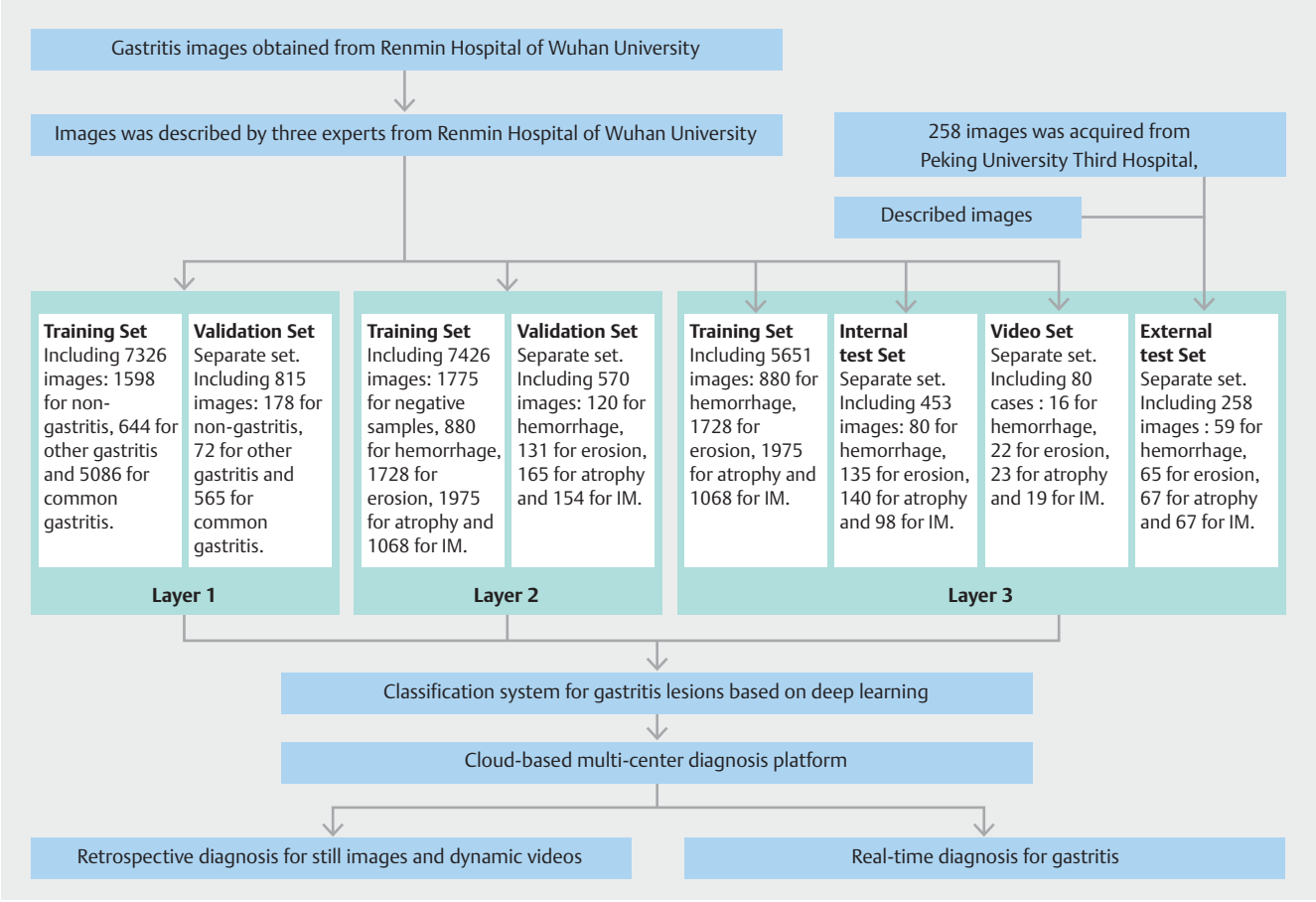
### Interobserver agreement of endoscopists

The kappa value for experts was higher than for non-experts with the three test sets. With the internal test set, experts reached substantial agreement in identifying AG/non-AG and erosion/ hemorrhage but moderate agreement in identifying GA/IM. Non-experts reached moderate agreement in most




▶ **Fig. 2** Representative images of four kinds of gastritis lesions. **a** Atrophy. **b** IM. **c** Erosion. **d** Hemorrhage. The first column are originals. The second are segmentation masks. The third shows: green dotted line and white box domain surrounds CNN predicting domain and blue dotted line and red box surrounds manual descripting domain. The fourth are demonstration: classification results and confidence are displayed.

cases. With the external test set, experts reached substantial agreement in most cases. Most non-experts reached moderate agreement but some of them reached substantial agreement. With the video set, experts reached perfect agreement in identifying CAG/non-AG and substantial agreement in identifying GA/IM and erosion/hemorrhage. Some endoscopists reached fair or moderate agreement with the video set.




▶ **Fig. 3** Flowchart.

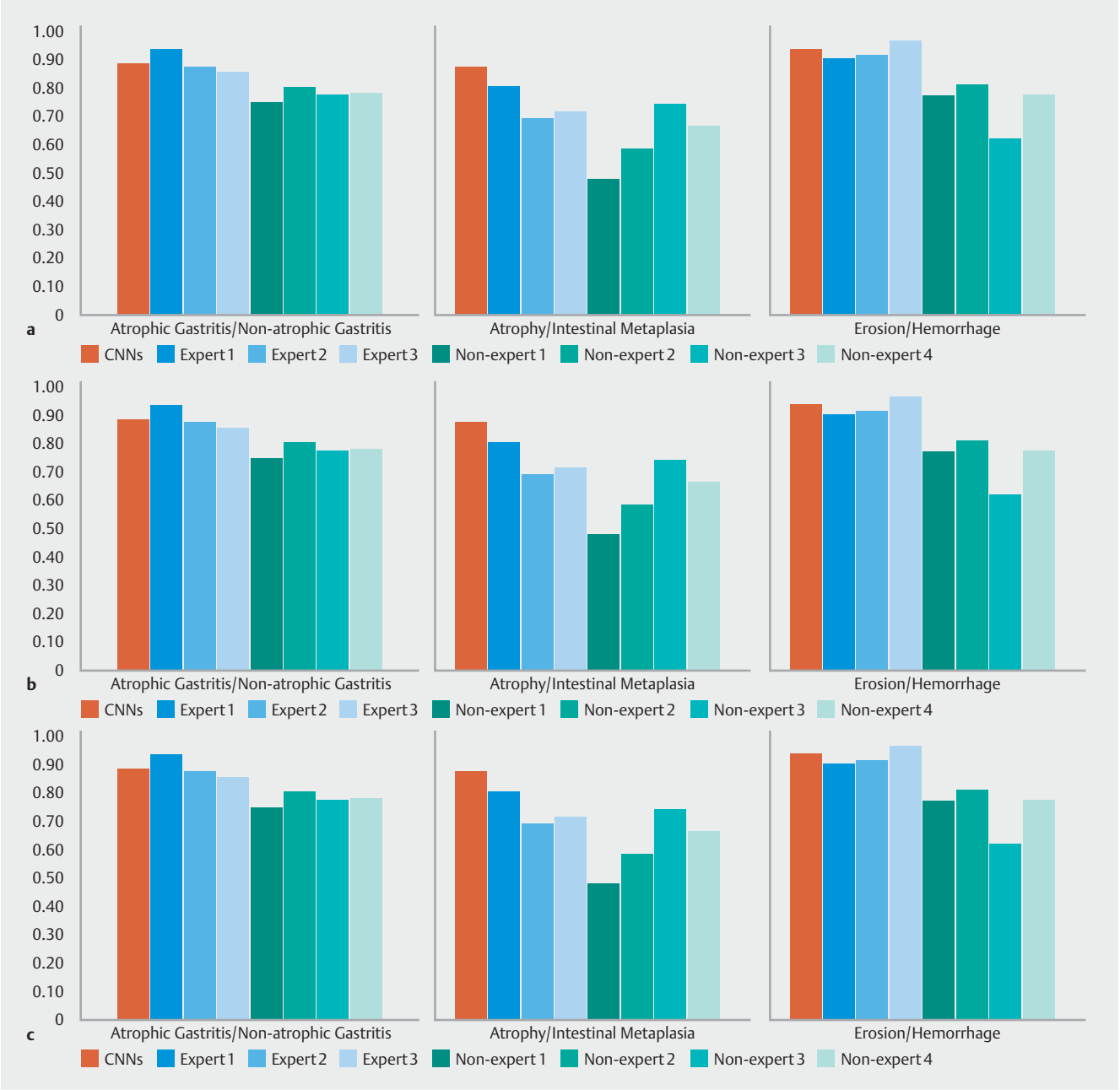
## Discussion

We constructed a DCNN-based gastritis lesion classification system, aiming to assist endoscopists in making an instant and precise diagnosis during examinations. The results previously described underscore the potential of our classification system. The performance of our model was better than endoscopists' average level, and even comparable to that of experts. Notably, unlike with the other test sets, we find no significant difference between CNNs and endoscopists with in video set. One reason is that the specificity of endoscopists in most classifications is on the high side, which is more than 80 % or 90 % in the video set, which suggests better performance on positive samples, especially those for which there were more images. Continuous image review may improve the accuracy because more images are available for analysis. The specificity of DCNN3 was associated with the proportion of IM images in the three test sets. A higher proportion of IM may lead to higher specificity, which means higher capability for distinguishing IM. Our system performed significantly better than the non-experts with the internal and external sets, but not significantly better with the video set. This is reasonable because lesions can be seen from different angles on video clips, providing endoscopists with more details. Non-experts had an uneven performance on identifying endoscopic lesions and the system may allow them and novices to make decisions more precisely.

Our system can tell if the patient has elevated-risk lesions such as GA and IM in the hotbed of carcinoma. In our previous work, WIESENCE (now named ENDOANGEL) was found to predict lesion location during examination. Our system can identify both the features and locations of the lesions. As reported, AG commonly affects the antrum, body, and fundus [31, 32]. Gastric cancer develops over a long period of progression [2]. Differentiated gastric cancer is associated with severe AG, and especially with IM [33]. Regular endoscopic examination is required in OLGA advanced stages [4]. Our system can examine stomachs comprehensively, ensuring that endoscopists do not miss lesions and allowing them to focus specifically on these hotbeds.

Outperforming endoscopists' average level, our system possessed the ability to alert endoscopists and reduce the rate of missed diagnoses. Owing to the decent grades that the system achieved with separate test sets, we believe it will have good stability in clinical practice. Moreover, it can function efficiently without fatigue.

Giving lesion details after examinations is another advantage of our model. With the scope close to the problem area, the system can help endoscopists determine, layer by layer,

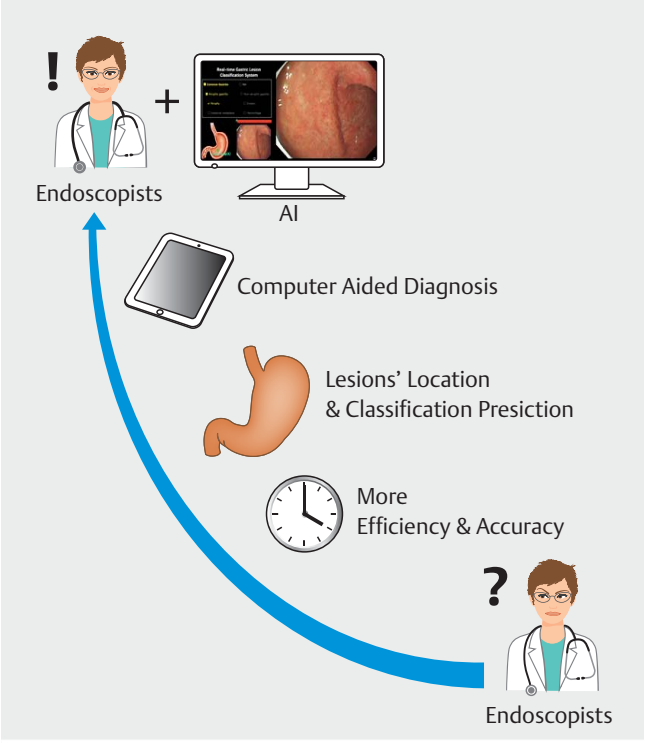



▶ **Fig. 4** Performance of endoscopists vs. model. **a**, **b**, **c** Accuracy of CNNs and endoscopists in the internal, external, and video test sets, respectively.

whether the patient has common gastritis, whether it is AG or non-AG, and what kind of lesions the patient has. All of those details also are summarized at the end of examinations. Endoscopists can make medical decisions more conveniently with prompting from the machine.

The system can play a role in training novices. After further improving the system, we believe it may be a powerful tool for training. For some hospitals that lack enough experts for teaching, our machine can help trainees with daily practice and also experts by facilitating spot testing. This would be an efficient method, costing less money and requiring fewer people than standard training.

The system proved to be applicable for clinical practice because it exhibited favorable results when used on the external test set. It can support endoscopist decision-making making by prompting for features of gastritis in lesions. This was the first study to describe location prediction-assisted gastritis lesion classification with a DCNN based system. Previous reports exist of DL with potential to recognize *H. pylori* infection and precancerous condition; however, those systems were constructed only for classification of *H. pylori*-relative gastritis or AG [8, 9, 20]. T. Itoh et al. [20] trained and tested their models using lesser curvature images, with the result that their clinical application is somewhat limited. Our system not only per-




▶ **Fig. 5** Summary chart.

formed lesions classification in real time but also labels the specific lesion location, making our research more clinically meaningful. With our system, endoscopists also will receive timely feedback for diagnosis. Moreover, the real-time photodocumentation with the system is convenient for doctors when writing endoscopy reports, which is time-saving.

There are still some limitations of our model worth improving. Primarily, segmentation helps remove unrelated background and leaves only lesions. This contributes to the precise diagnosis performed by the system. Residual interference may still influence the results. Diverse lesions may influence the purity of training sets, but we have a reasonable control for that. Moreover, we only used white-light images in this study. The setting used for testing was community hospitals. Our system cannot prompt endoscopists to perform biopsies, but our team is investigating that functionality in ongoing research. In this study, we used video clips for testing, but classification of gastritis lesions in short videos is more in line with clinical practice, because endoscopists usually do not spend a long time observing benign lesions. This was a retrospective study, and as such, selection bias was unavoidable; a prospective study is being planned to further validate real-time use of AI in clinical practice. The prevalence of *H. pylori*-negative gastric cancer has recently increased, and our system is not able to predict risk factors for it very well. We are collecting related cases to complete our system [34].

The results with the three test sets prove our model's stability. We believe it will perform similarly well in clinical practice. As an accuracy-volume curve shows (**Supplementary Fig. 3**), the accuracy of DCNNs improves with increasing image volume, and we can improve our system by add more typical images. We are preparing to conduct clinical trials with the system. Another supplementary experiment is in the works, focusing on the Kimura-Takemoto Classification for gastric cancer risk assessment [10] Nakahira, H. et al. has reported on significant work with a gastric cancer risk stratification AI system. They divided the images into four groups according to their locations. [21] Our system predicts the locations in more details. We believe it will be a powerful tool for gastric cancer risk assessment.

## Conclusions

This study provides proof of the that an auxiliary diagnostic system with DL can be used to established the classification of gastritis lesions, summarize their relevant characteristics, and prompt endoscopists about the findings. The accuracy of the models in test sets was comparable to that of expert endoscopists. Nevertheless, the system still requires further improvement to achieve the goal of clinically summarizing gastritis lesions using AI (▶ **Fig. 5**).

## Acknowledgements

The authors thjank all these who helped them during this study. Their deepest gratitude goes first and foremost to Professor Honggang Yu and Shigang Ding, whose thoughts and constructive suggestions played a crucial role in this study. Xiao Hu and Shan Hu contributed to the development of an algorithm that ensured the quality of the classification system. Renquan Luo, Jing Wang, and Chao Li helped collected data for our work. Ganggang Mu, Hongyan Li, and Jixiang Zhang arrived at consensus on the image classification as reviewers. Yanxia Li and Zhanyue Niu participated in testing the model and assisted with analysis of the results. Lianlian Wu managed the collection and analysis of the data, which was a key step in our study. Yijie Zhu wrote the manuscript. We also thank Chao Huang, Fei Liao, Pengbo Wu, Ya Liu, Zhengqiang Wang, Zihua Lu, Xi Chen, Zehua Dong, Yunchao Deng, Jing Zhang, Yan Xue, and Jun Zhang, whose gracious help facilitated completion of our study.

### Competing interests

Drs. Shan Hu, Xiao Hu, and Chao Li are research staff members of Wuhan EndoAngel Medical Technology Company.

### References

[1] Bray F, Ferlay J, Soerjomataram I et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018; 68: 394–424

[2] Banks M, Graham D, Jansen M et al. British Society of Gastroenterology guidelines on the diagnosis and management of patients at risk of gastric adenocarcinoma. Gut 2019; 68: 1545–1575

[3] Correa P. Human Gastric Carcinogenesis: A Multistep and Multifactorial Process—First American Cancer Society Award Lecture on Cancer Epidemiology and Prevention. Cancer Research 1992; 52: 6735

[4] Rugge M, Meggio A, Pravadelli C et al. Gastritis staging in the endoscopic follow-up for the secondary prevention of gastric cancer: a 5-year prospective study of 1755 patients. Gut 2019; 68: 11–17

[5] Leung WK, Ho HJ, Lin J-T et al. Prior gastroscopy and mortality in patients with gastric cancer: a matched retrospective cohort study. Gastrointest Endosc 2018; 87: 119–127.e113

[6] 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: 365–388

[7] Du Y, Bai Y, Xie P et al. Chronic gastritis in China: a national multicenter survey. BMC Gastroenterology 2014; 14: 21

[8] Guimaraes P, Keller A, Fehlmann T et al. Deep-learning based detection of gastric precancerous conditions. Gut 2020; 69: 4–6

[9] Zhang Y, Li F, Yuan F et al. Diagnosing chronic atrophic gastritis by gastroscopy using artificial intelligence. Dig Liver Dis 2020; 52: 566–572

[10] Kimura K, Takemoto T. An Endoscopic Recognition of the Atrophic Border and its Significance in Chronic Gastritis. Endoscopy 1969; 1: 87–97. doi:10.1055/s-0028-1098086

[11] Rugge M, Meggio A, Pennelli G et al. Gastritis staging in clinical practice: the OLGA staging system. Gut 2007; 56: 631–636

[12] Sugano K, Tack J, Kuipers EJ et al. Kyoto global consensus report on Helicobacter pylori gastritis. Gut 2015; 64: 1353–1367

[13] TYTGAT GNJ. Endoscopic appearances in gastritis/duodenitis. The Sydney System: Endoscopic division 1991; 6: 223–234

[14] Jin EH, Chung SJ, Lim JH et al. Training Effect on the inter-observer agreement in endoscopic diagnosis and grading of atrophic gastritis according to level of endoscopic experience. J Korean Med Sci 2018; 33: e117

[15] Ono S, Dohi O, Yagi N et al. Accuracies of endoscopic diagnosis of helicobacter pylori-gastritis: multicenter prospective study using white light imaging and linked color imaging. Digestion 2020; 101: 624–630

[16] Dutta AK, Sajith KG, Pulimood AB et al. Narrow band imaging versus white light gastroscopy in detecting potentially premalignant gastric lesions: a randomized prospective crossover study. Ind J Gastroenterol 2013; 32: 37–42

[17] Esteva A, Kuprel B, Novoa RA et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017; 542: 115–118

[18] 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: 370–378

[19] Togo R, Yamamichi N, Mabe K et al. Detection of gastritis by a deep convolutional neural network from double-contrast upper gastrointestinal barium X-ray radiography. J Gastroenterol 2019; 54: 321–329

[20] Itoh T, Kawahira H, Nakashima H et al. Deep learning analyzes Helicobacter pylori infection by upper gastrointestinal endoscopy images. Endosc Int Open 2018; 6: E139–E144

[21] Nakahira H, Ishihara R, Aoyama K et al. Stratification of gastric cancer risk using a deep neural network. 2020; 4: 466–471

[22] Wu L, Zhang J, Zhou W et al. Randomised controlled trial of WISENSE, a real-time quality improving system for monitoring blind spots during esophagogastroduodenoscopy. Gut 2019; 68: 2161–2169

[23] Wu L, Zhou W, Wan X et al. A deep neural network improves endoscopic detection of early gastric cancer without blind spots. Endoscopy 2019; 51: 522–531

[24] Chen D, Wu L, Li Y et al. Comparing blind spots of unsedated ultrafine, sedated, and unsedated conventional gastroscopy with and without artificial intelligence: a prospective, single-blind, 3-parallel-group, randomized, single-center trial. Gastrointest Endosc 2020; 91: 332–339.e333

[25] Ruiz B, Garay J, Correa P et al. Morphometric evaluation of gastric antral atrophy: improvement after cure of Helicobacter pylori infection. Am J Gastroenterol 2001; 96: 3281–3287

[26] Genta RM. Recognizing atrophy: another step toward a classification of gastritis. Am J Surg Path 1996; 20: S23–S30

[27] Rugge M, Genta RM. Staging and grading of chronic gastritis. Hum Pathol 2005; 36: 228–233

[28] He K, Zhang X, Ren S et al. Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision & Pattern Recognition; 2016

[29] Zhou Z, Siddiquee MMR, Tajbakhsh N et al. UNet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Transact Med Imaging 2019: doi:10.1109/tmi.2019.2959609

[30] Shao L, Zhu F, Li X. Transfer learning for visual categorization: a survey. IEEE Transact Neural Net Learning Sys 2015; 26: 1019–1034

[31] You WC, Li JY, Blot WJ et al. Evolution of precancerous lesions in a rural Chinese population at high risk of gastric cancer. Int J Cancer 1999; 83: 615–619

[32] Blaser MJ. Type B gastritis, aging, and Campylobacter pylori. Arch Int Med 1988; 148: 1021–1022

[33] Uemura N, Okamoto S, Yamamoto S et al. Helicobacter pylori infection and the development of gastric cancer. N Engl J Med 2001; 345: 784–789

[34] Yamamoto Y, Fujisaki J, Omae M et al. Helicobacter pylori-negative gastric cancer: characteristics and endoscopic findings. Dig Endosc 2015; 27: 551–561