Identification of upper GI diseases during screening gastroscopy using a deep convolutional neural network algorithm

Hang Yang, BA,[1,*] Yu Wu, PhD,[2,*] Bo Yang, MA,[3,*] Min Wu, PhD,[4,*] Jun Zhou, MS,[1] Qin Liu, BA,[1] Yifei Lin, PhD,[5] Shilin Li, MS,[1] Xue Li, MS,[1] Jie Zhang, BA,[1] Rui Wang, PhD,[1] Qianrong Xie, BA,[1] Jingqi Li, PhD,[6] Yue Luo, BA,[7] Mengjie Tu, BA,[1,8] Xiao Wang, MS,[3] Haitao Lan, PhD,[9] Xuesong Bai, MD,[3] Huaping Wu, PhD,[10] Fanwei Zeng, PhD,[11] Hong Zhao, PhD,[12] Zhang Yi, PhD,[2] Fanxin Zeng, PhD[1,2]

Sichuan, Chengdu, Heilongjiang, Guangdong, Beijing, China

Background and Aims: The clinical application of GI endoscopy for the diagnosis of multiple diseases using artificial intelligence (AI) has been limited by its high false-positive rates. There is an unmet need to develop a GI endoscopy AI-assisted diagnosis system (GEADS) to improve diagnostic accuracy and clinical utility.

Methods: In this retrospective, multicenter study, a convolutional neural network was trained to assess upper GI diseases based on 26,228 endoscopic images from Dazhou Central Hospital that were randomly assigned (3:1:1) to a training dataset, validation dataset, and test dataset, respectively. To validate the model, 6 external independent datasets comprising 51,372 images of upper GI diseases were collected. In addition, 1 prospective dataset comprising 27,975 images was collected. The performance of GEADS was compared with endoscopists with 2 professional degrees of expertise: expert and novice. Eight endoscopists were in the expert group with >5 years of experience, whereas 3 endoscopists were in the novice group with 1 to 5 years of experience.

Results: The GEADS model achieved an accuracy of .918 (95% confidence interval [CI], .914-.922), with an F1 score of .884 (95% CI, .879-.889), recall of .873 (95% CI, .868-.878), and precision of .890 (95% CI, .885-.895) in the internal validation dataset. In the external validation datasets and 1 prospective validation dataset, the diagnostic accuracy of the GEADS ranged from .841 (95% CI, .834-.848) to .949 (95% CI, .935-.963). With the help of the GEADS, the diagnosing accuracies of novice and expert endoscopists were significantly improved ($P < .001$).

Conclusions: The AI system can assist endoscopists in improving the accuracy of diagnosing upper GI diseases. (Gastrointest Endosc 2022;96:787-95.)

(footnotes appear on last page of article)

Upper GI diseases, including esophageal cancer, gastric cancer, and ulcers, are very common in the general population, but inappropriate diagnoses often result in serious public health burdens and impaired quality of life.[1-3] Because the symptoms or endoscopic findings are complex and diverse, patients with different diagnoses may need individualized treatments. Therefore, an accurate diagnosis could not only help improve the prognosis of patients and reduce medical costs, but could also play an essential role in precision medicine.

White-light endoscopy is the primary noninvasive method in the diagnosis of the upper GI diseases.[4-7] However, the enormous amounts of images produced by endoscopy can bring a heavy workload to clinicians and even cause eye fatigue, leading to a certain rate of missed diagnoses.[8] To overcome such challenges, multifarious endoscopic strategies such as narrow-band imaging and chromoendoscopy have been initiated and developed and have significantly increased the diagnostic accuracy of GI diseases.[9-13] However, physician experience can also affect the diagnostic accuracy for upper GI diseases.

Recently, artificial intelligence (AI) has been widely used to detect early gastric cancer,[12,14,15] chronic atrophic gastritis,[16,17] and *Helicobacter pylori* infection.[18-20] Many encouraging studies have advanced the application of AI in endoscopic imaging.[17,21-23] Thus, use of AI for upper GI diseases was believed to help bridge the gap between the differences in diagnosis by endoscopists and make the examination results more objective and standardized. Nevertheless, most studies were single-center and used binary classification alone, which makes it difficult to meet the clinical demand.[24-26]

Here, we developed a high-performance multiclassification GI endoscopy AI-assisted diagnosis system (GEADS) consisting of StoNet, which was developed for the diagnosis of upper GI diseases, and an anatomic localization model using a high volume of endoscopic images from multiple centers to achieve a high diagnostic accuracy of upper GI diseases and lesion location. A prospective study was conducted to evaluate the effectiveness of the AI model in clinical practice.

METHODS

Study design and participants

This multicenter study was performed in 7 hospitals in China: Dazhou Central Hospital, Xuanhan People's Hospital, Quxian People's Hospital, Kaijiang People's Hospital, Tongchuan People's Hospital, Dachuan People's Hospital, and Dazhu People's Hospital. We collected endoscopic images from Dazhou Central Hospital to develop the GEADS to identify upper GI diseases and externally verified the GEADS using endoscopic images from 6 other hospitals.

The monitor for verifying the performance of the GEADS was connected to the original endoscopic monitor, and an independent dataset of consecutive participants who received upper GI endoscopy was enrolled. These participants were grouped into the prospective validation dataset.

We retrospectively searched the medical records from all participating hospitals and mainly included participants aged >18 years. Five groups of people were included: patients with cancer, erosion, polyp, and ulcers and normal participants. All cancer images were obtained from patients with a definitive pathologic diagnosis. Patients with incomplete information, undetermined pathologic diagnosis of cancer, gastrectomy history, reflux esophagitis, *Helicobacter pylori* infection, Barrett's esophagus, diverticulum, chronic atrophic gastritis, esophageal varices, gastric varicose veins, gastric antrum white spot, surgical history of esophageal cancer, carditis, massive GI bleeding, or without white-light endoscopic images were excluded. When we screened images from datasets, we excluded overlapping images of every patient, including the internal training, validation, test, and prospective datasets and external validation datasets.

This study was approved by the Institutional Review Board of Dazhou Central Hospital and performed according to the Declaration of Helsinki. Each participating hospital was exempt from informed consent by the institutional review board because of the retrospective nature of the study. Participants in the prospective dataset were informed of the purpose of the study and provided written informed consent (IRB00000012-21001).

Endoscopy and image quality control

The images in this study were acquired using different endoscopes (Evis Lucera CLV-260, Evis Lucera CLV-290 [Olympus Medical Systems, Tokyo, Japan], and Fujinon EG-99WR [Fujifilm Medical Systems, Tokyo, Japan]). All upper GI endoscopic images were converted to JPEG format. The collected endoscopic images were cleaned by 5 endoscopists with 3 years of experience from Dazhou Central Hospital according to the same standard, and some images of poor quality such as blurred, large, and bright spots; defocus; foam; mucus; many interference bands; and low picture brightness and those with food residue were deleted. The images were reviewed, classified, and framed by 2 endoscopists with less than 5 years of experience. Ground truth was established based on a consensus from an independent group of 4 endoscopists with 6 or more years of experience. For these nonconsensus images, we invited a senior endoscopist to judge.

Development of the StoNet model

First, we developed a StoNet model based on our large number of clinical endoscopic images to identify upper GI diseases (Fig. 1A). The endoscopic images from Dazhou Central Hospital were randomly assigned (3:1:1) to the training, validation, and test datasets, respectively, for developing the StoNet. Internal validation datasets (validation and test datasets) were used to evaluate the performance of the StoNet.

The pattern of StoNet is shown in Supplementary Figure 1 (available online at www.giejournal.org). The training procedure was completed after 300 epochs on the training dataset. The training procedure of the accuracy and cross-entropy loss was finished (Supplementary Fig. 2, available online at www.giejournal.org). The best trainable weight for StoNet was retained by the validation dataset in limited epochs.

StoNet was based on the concept of DenseNet121 (2017 IEEE Conference on Computer Vision and Pattern Recognition [CVPR], Honolulu, Haw, USA) and the feature pyramid network (Supplementary Fig. 1). The input of the model was the white-light endoscopic images. The output was a standard 5-classification task to determine whether the input images contained a specific type (normal, cancer, erosion, polyp, or ulcer). The proposed model was trained with the training dataset using a back-propagation algorithm, and the parameters were then fine-tuned using the transfer learning technique. See Supplementary Methods (available online at www.giejournal.org) for additional details regarding the proposed network. In addition, several typical convolutional neural network models such as DenseNet121, ResNet50 (2016 IEEE Conference on Computer Vision and Pattern Recognition [CVPR], Las Vegas, Nev, USA), and VGG-11 (VGG-11: 2015 ICLR, San Diego, Calif, USA) were also used to compare the difference between the typical convolutional neural network and StoNet.

VGG-11 used small filters because of fewer parameters and stacked more of them instead of having larger filters. It has the same effective receptive field as if only one 7×7 convolutional layer is used. ResNet50 was a jump

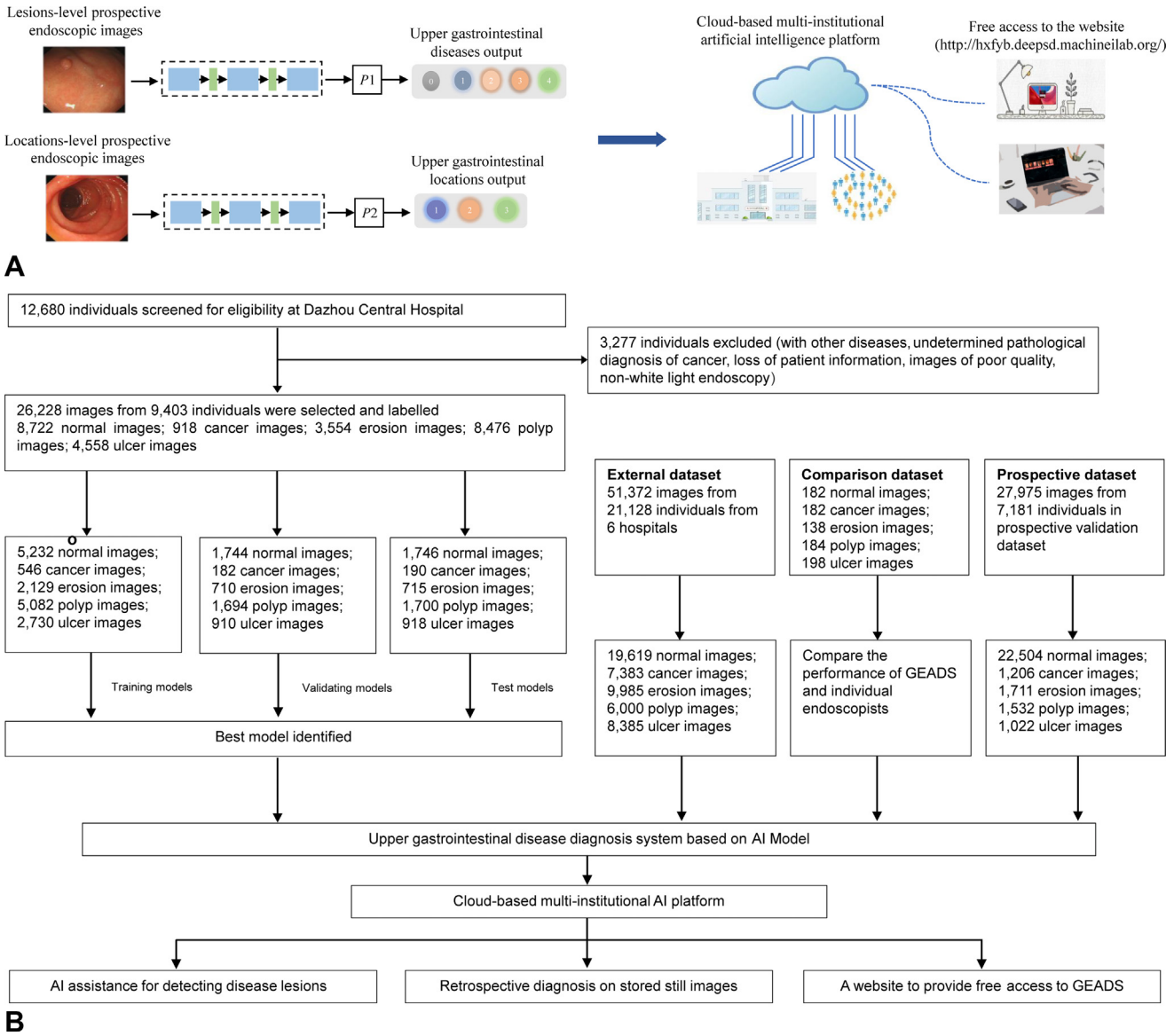


Figure 1. Trial flow diagram. **A,** Trial design. **B,** Screening and study flow. *AI*, Artificial intelligence; *GEADS*, GI endoscopy artificial intelligence–assisted diagnosis system.

connection between each layer and a previous layer (generally 2-3 layers). The connection method is through element level addition. In DenseNet121, each layer will be concatenated with all previous layers in the channel dimension and used as the input of the next layer. Compared with the classifier of general neural network, which directly depends on the characteristics of the last layer of the network (with the highest complexity), DenseNet121 can make comprehensive use of the characteristics with low shallow complexity, so it is easier to get a smooth decision function with better generalization performance.

Development of the anatomic localization model

At the same time, we used a total of 4356 endoscopic images, including 1482 esophagus, 1769 stomach, and 1105 duodenum images, to develop a model for identifying anatomic location (including esophagus, stomach, and duodenum) (Supplementary Table 1, available online at www.giejournal.org). It was divided into the training, validation, and test datasets at a ratio of 3:1:1, respectively. The diagnostic accuracy, precision, recall, and F1 score of VGG-11, ResNet50, and DenseNet121 were calculated to compare performance.

Validation of the GEADS

GEADS consisted of StoNet and the anatomic localization model. We validated the performance of StoNet in identifying 5 classifications of upper GI diseases in patients using the 6 external datasets, comparison dataset, and prospective dataset (Fig. 1B). We then evaluated the diagnostic performance of the anatomic localization model in discriminating esophagus,

stomach, and duodenum using VGG-11, ResNet50, and DenseNet121. We built a website that provides free consulting services for patients and clinicians after uploading their endoscopic images on the website platform (Fig. 1A).

In the comparison dataset, 2 groups of 11 endoscopists with varying degrees of expertise (expert, ≥5 years of experience; novice, 1-5 years of experience), who were unaware of the patients' clinical information, were asked to independently complete the diagnosis of upper GI diseases with the same dataset and record the experimental results. The same endoscopists then conducted the same testing experiments with the AI-assisted system after 2 weeks to compare the diagnostic accuracy of endoscopists with and without the AI-assisted system. None of the 11 endoscopists was involved in the screening and labeling of the 884 images, which were also mixed up and hidden before the endoscopists' assessments.

Statistical analysis

The diagnostic accuracy, F1 score, recall, and precision for the GEADS identification of the 5 types of endoscopic images were evaluated by calculating the 95% confidence interval (CI). The F1 score is the weighted average of precision and recall, which is typically more useful than accuracy when it comes to an uneven class distribution. A receiver operating characteristic curve was used to demonstrate the diagnostic ability of our proposed model (StoNet) in categorizing patients into 5 classes. A larger area under the curve (AUC) indicated better diagnostic performance. We also used the confusion matrix to provide a more detailed analysis than a mere proportion of correct classifications, where each row of the matrix represents the instances in a predicted class and each column represents the instances in an actual class. Statistical analyses were performed using Python software (version 3.0, Python Software Foundation, Wilmington, Del, USA). The differences between the GEADS and endoscopists in the accuracy were compared using a 2-tailed unpaired Student *t* test.

RESULTS

Clinical characteristics

Between June 28, 2017 and October 31, 2020, 26,228 images from 9403 participants were obtained from the upper GI endoscopic imaging database at Dazhou Central Hospital (Fig. 1B). In the training dataset, we used 15,719 upper GI images from the Dazhou Central Hospital (Fig. 1B). We found that in the training dataset 27.8% were normal, whereas 2.8% had cancer, 14.3% had erosion, 38.1% had polyps, and 17.0% had ulcers (Table 1). In the prospective validation dataset, 27,975 images from 7181 individuals were prospectively collected and labeled (Fig. 1B). The image data obtained from the 6 other participating hospitals between January 1, 2018 and October 31, 2020 is summarized in Table 1. Overall, 51,372 endoscopic images from 21,128 individuals were

TABLE 1. Baseline characteristics

	Dazhou Central Hospital validation				External validation					
	Training	Validation	Test	Prospective	Xuanhan People's Hospital	Quxian People's Hospital	Kaijiang People's Hospital	Tongchuan People's Hospital	Dachuan People's Hospital	Dazhu People's Hospital
Sex										
Female	3314 (58.3)	1055 (55.6)	1054 (58.1)	3898 (54.3)	1202 (53.0)	918 (39.2)	2117 (50.4)	855 (55.8)	1864 (51.0)	3546 (49.7)
Male	2375 (41.7)	844 (44.4)	761 (41.9)	3283 (45.7)	1064 (47.0)	1422 (60.8)	2081 (49.6)	678 (44.2)	1792 (49.0)	3589 (50.3)
Mean age, y (standard deviation)	53.34 (13.41)	53.44 (13.43)	53.19 (13.54)	52.40 (14.49)	51.31 (14.39)	56.20 (13.68)	53.38 (14.75)	55.76 (35.19)	55.33 (13.43)	54.61 (13.41)
18-42 y	979 (17.2)	317 (16.7)	318 (17.5)	1479 (20.6)	519 (22.9)	332 (14.2)	760 (18.1)	293 (19.1)	545 (14.9)	1013 (14.2)
43-63 y	3271 (57.5)	1107 (58.3)	1038 (57.2)	3943 (54.9)	1237 (54.6)	1177 (50.3)	2351 (56.0)	739 (48.2)	2055 (56.2)	4,046 (56.7)
≥64 y	1439 (25.3)	475 (25.0)	459 (25.3)	1759 (24.5)	510 (22.5)	831 (35.5)	1087 (25.9)	501 (32.7)	1056 (28.9)	2076 (29.1)
Normal	1582 (27.8)	531 (28.0)	511 (28.2)	2767 (38.5)	800 (35.3)	682 (29.1)	2164 (51.5)	565 (36.9)	1284 (35.1)	1848 (25.9)
Cancer	162 (2.8)	56 (2.9)	62 (3.4)	181 (2.5)	51 (2.3)	374 (16.0)	164 (3.9)	53 (3.5)	354 (9.7)	438 (6.1)
Erosion	812 (14.3)	267 (14.1)	254 (14.0)	1990 (27.7)	649 (28.6)	320 (13.7)	1066 (25.4)	376 (24.5)	570 (15.6)	1194 (16.7)
Polyp	2165 (38.1)	724 (38.1)	675 (37.2)	923 (12.9)	398 (17.6)	402 (17.2)	158 (3.8)	260 (17.0)	655 (17.9)	1533 (21.5)
Ulcer	968 (17.0)	321 (16.9)	313 (17.2)	1320 (18.4)	368 (16.2)	562 (24.0)	646 (15.4)	279 (18.2)	793 (21.7)	2122 (29.8)

Values are n (%) unless otherwise defined.

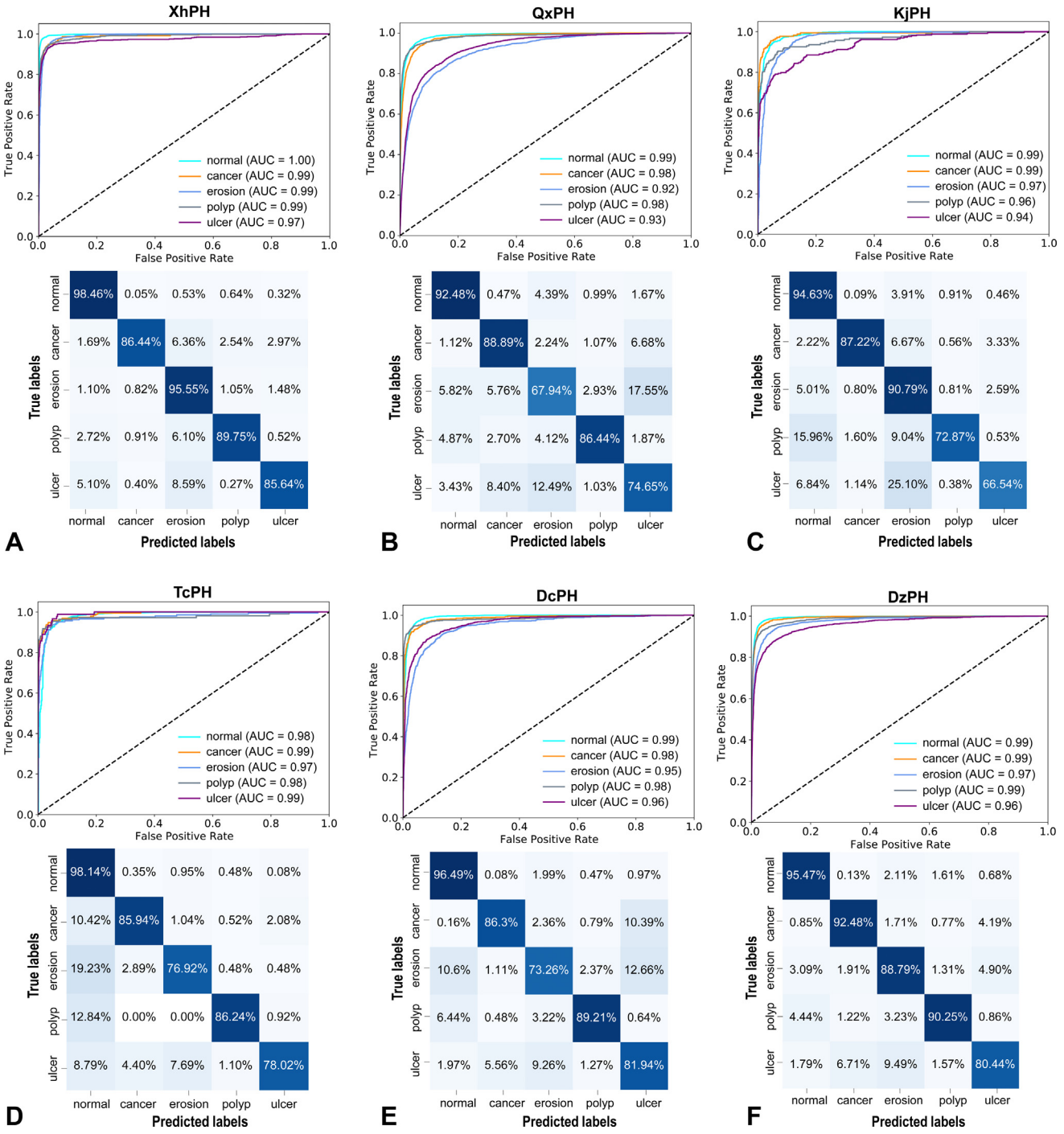


Figure 2. Receiver operating characteristic curves and confusion matrix illustrate the ability of the GI endoscopy artificial intelligence–assisted diagnosis system (GEADS) to detect upper GI diseases. **A,** Receiver operating characteristic curves and confusion matrix of the Xuanhan People's Hospital in the validation dataset. **B,** Receiver operating characteristic curves and confusion matrix of the Quxian People's Hospital in the validation dataset. **C,** Receiver operating characteristic curves and confusion matrix of the Kaijiang People's Hospital in the validation dataset. **D,** Receiver operating characteristic curves and confusion matrix of the Tongchuan People's Hospital in the validation dataset. **E,** Receiver operating characteristic curves and confusion matrix of the Dachuan People's Hospital in the validation dataset. **F,** Receiver operating characteristic curves and confusion matrix of the Dazhu People's Hospital in the validation dataset. *AUC*, Area under the curve.

used to validate GEADS. In addition, 1895 patients with a pathologically confirmed report were analyzed, and their cancer information is presented in Supplementary Table 2 (available online at www.giejournal.org).

Performance of GEADS model and multicenter validation

The diagnostic accuracies were .918 (95% CI, .914-.922) and .916 (95% CI, .910-.920) for the internal validation and

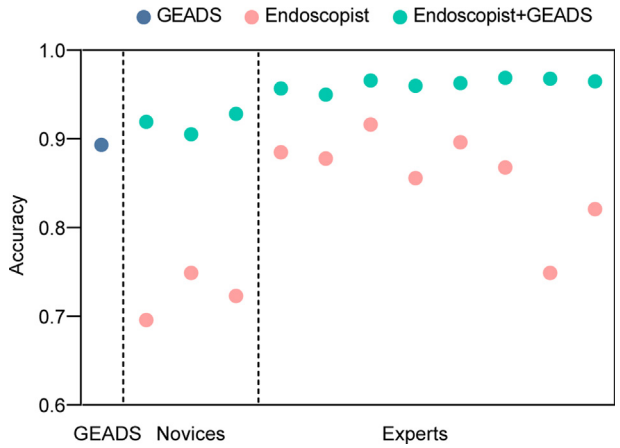


Figure 3. Comparison of diagnostic performance between the artificial intelligence system and endoscopists. *GEADS*, GI endoscopy artificial intelligence–assisted diagnosis system.

test datasets, respectively. The detailed classification results of the GEADS in the internal validation datasets are shown in the Supplementary Figure 3 (available online at www.giejournal.org). In the prospective dataset, the GEADS showed better performance (accuracy, .925; 95% CI, .922-.929) (Supplementary Table 3, available online at www.giejournal.org). Similarly, high diagnostic accuracies were observed by GEADS in the 6 external validation datasets (ranging from .841 [95% CI, .834-.848] to .949 [95% CI, .935-.963]) (Supplementary Table 3). The F1 score, recall, and precision were higher than .800 in the internal validation dataset and 6 external validation datasets. In the first external validation dataset (Xuanhan People's Hospital), our GEADS achieved a perfect AUC of 1.000 for normal versus all other groups together and an AUC of .990 for the diagnosis of cancer, erosion, and polyp (Fig. 2A). The AUC achieved by our GEADS in the second to the sixth external validation datasets are summarized in Figure 2B to F. Overall, the AUCs of the 6 external validation datasets were high, ranging from .930 to 1.000 (Fig. 2). The GEADS was accurate in identifying patients from upper GI images (normal, cancer, erosion, polyp, and ulcer) in all 7 validation datasets (Supplementary Table 3). The detailed classification of the GEADS performance regarding the correlation of the predicted labels is described as a confusion matrix (Fig. 2).

The StoNet, the core algorithm of our GEADS, performed with high accuracy in identifying normal and upper GI diseases. Diagnostic accuracies were .965 (95% CI, .935-.963) in discriminating normal images, .987 (95% CI, .984-.990) for cancer, .960 (95% CI, .955-.965) for erosion, .969 (95% CI, .965-.974) for polyp, and .950 (95% CI, .944-.956) for ulcer. In addition, the StoNet model had superior precision, recall, and F1 score compared with VGG-11, ResNet50, and DenseNet121 for the diagnosis of upper GI diseases (Supplementary Table 4, available online at www.giejournal.org). Taken together, these results demonstrated that StoNet performs better in a multiclass diagnosis. Meanwhile, the DenseNet121 algorithm was used to identify anatomic location and achieved good diagnostic performance, superior to VGG-11 and similar to ResNet (Supplementary Fig. 4 and Supplementary Table 5, available online at www.giejournal.org).

Comparison of the GEADS and endoscopists

An independent comparison dataset of 884 endoscopic images, including 182 normal, 182 cancer, 138 erosion, 184 polyp, and 198 ulcer images, was used to compare the performance of the GEADS with that of endoscopists in classifying upper GI diseases (Fig. 1B). Eleven endoscopists were divided into 2 groups: 3 endoscopists in the novice group with 1 to 5 years of experience and 8 endoscopists in the expert group with >5 years of experience. The GEADS yielded a false-positive rate of 2.7% and a false-negative rate of 11.0% compared with a mean of 6.3% and 27.1%, respectively, by the novices (Supplementary Table 6, available online at www.giejournal.org).

We used predicted labels and true labels to create a matrix to evaluate and compare the performance of the GEADS and endoscopists (Supplementary Fig. 5, available online at www.giejournal.org). After our statistical analysis regarding accuracy, the GEADS performed as well as expert endoscopists ($P = .557$), but the performance of the

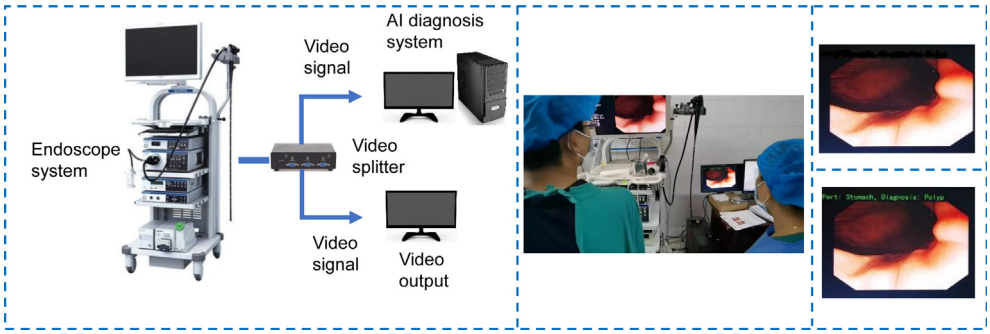


Figure 4. Use of the GEADS to identify lesions during endoscopic examination. The computer on which the GEADS was installed was connected directly to an endoscopy unit, allowing fully automated diagnosis during endoscopic examination. *AI*, Artificial intelligence; *GEADS*, GI endoscopy artificial intelligence–assisted diagnosis system.

GEADS was significantly better than novice endoscopists ($P = .034$) (Fig. 3 and Supplementary Table 6).

To investigate whether the GEADS could help endoscopists improve their diagnostic performance, each endoscopist was asked to make diagnoses with the assistance from the GEADS results. The follow-up GEADS-assisted diagnostic testing by endoscopists was performed 2 weeks after the initial test. With the help of the GEADS, the accuracy of novices and experts in diagnosing upper GI diseases was significantly improved ($P < .001$) (Fig. 3 and Supplementary Table 6). We used GEADS to identify lesions during endoscopic examination (Fig. 4). We summarized the misclassified images of physicians, the GEADS, and AI-assisted physicians (Supplementary Fig. 6, available online at www.giejournal.org).

DISCUSSION

In this study, we developed a multiclass deep learning model, the GEADS, for identifying upper GI diseases using 106,459 endoscopy images from 7 hospitals. The GEADS showed excellent accuracy, sensitivity, and specificity in diagnosing and targeting upper GI diseases using retrospective images. Moreover, the accuracy of novices and experts in diagnosing upper GI diseases was significantly improved with the assistance of the GEADS.

In terms of methodology, the novel StoNet algorithm was proposed based on the DenseNet121 framework to distinguish different types of upper GI diseases. The StoNet algorithm would benefit from combining high-dimensional features with low-dimensional features, which would allow the GEADS to learn more features of the endoscopic images, thus improving the accuracy of the model. In addition, we performed a test on the same dataset for the diagnosis of disease to compare with other traditional convolutional neural network methods, such as VGG-11, ResNet50, and DenseNet121.[15,16,27,28] Ultimately, we found that the StoNet algorithm achieved better performance with an accuracy of .965.

In clinical practice, basic diagnosis reports should consist of a description of the upper GI diseases and their anatomic location. For disease description, however, previous studies mainly considered partial or binary classifications, which might cause missed or delayed diagnosis of upper GI diseases.[12,15,29,30] The current automatic recognition model with relatively high accuracy has been applied in the diagnosis of gastric tumors but still faces difficulties in recognizing other lesions such as polyps or ulcers. Therefore, the GEADS, with its multiclassification model, has the additional advantage of distinguishing a variety of other noncancer diseases, including erosion, polyps, and ulcers.

Furthermore, we trained the GEADS to identify anatomic locations and imitate clinicians' diagnostic processes for upper GI diseases. the GEADS divided the structure of the upper GI tract into 3 sections: esophagus, stomach, and duodenum. This design is helpful in understanding the location reflected in the images.

A gastroscope generally finds lesions in the mucosal layer, which cannot exclude whether cancer cells are invading into the deeper part. The GEADS was trained on cancer images with pathologic results and found some features related to submucosal lesions that were often ignored by doctors. For some hidden small polyps, endoscopists may miss them and thus the diagnosis, but the GEADS can find this kind of polyp exactly. However, the GEADS did identify some light spots and foam parts as polyps, ulcers, or erosion because it relies on color, texture, and background clues to identify what it sees. With the aid of the GEADS, the accuracy of recognition was improved, but some endoscopists made judgments based on their own experience and ignored the correct tips from AI.

Endoscopy is a widely performed technique; however, the learning curve for endoscopists is steep. Because nonexperts might misdiagnose upper GI diseases,[21,29,31] numerous tools have been invented to cope with this situation. The GEADS could not only achieve expert-level accuracy but also significantly improve the performance of novices. It is hoped that the GEADS can alleviate the severe imbalance of health resources in underdeveloped countries and regions.

In recent years, the application of deep learning in digestive endoscopy has made great progress. He et al[32] trained a convolutional neural network on 4667 images of magnifying image-enhanced endoscopy for diagnosing early gastric cancer. The model achieved an accuracy of 88.44% and 90.4% in the internal and external datasets, respectively. A large number of videos were trained and validated, which achieved an excellent accuracy, but there was a lack of research on a variety of diseases, such as ulcer and polyp. Wu et al[7] trained and tested a deep convolution neural network for identifying early gastric cancer using 9151 images with an accuracy of .925, and in unprocessed real-time EGD videos, the deep convolution neural network enables the automatic detection of early gastric cancer and monitoring blind spots. However, this study had a binary classification and single-center dataset, limiting exceptional types.

The GEADS was validated in 7 different hospitals. In addition, we included various types of images and achieved multiclassification of diseases. Despite the improvement in diagnostic performance, the GEADS can inevitably lead to misdiagnoses. The false-positive and false-negative rates of the GEADS were 2.7% and 11%, respectively. However, both the false-positive and false-negative rates of the GEADS were lower than those of novices and close to experts. The main reasons for the false rate by GEADS might be light spots on endoscopy images or other fusion of multiple lesions, such as cancer with a round bulging growth, cancer with a characterized ulcer, and an ulcer or erosion on a polyp.

This study had some limitations. First, in the process of cleaning the data, we excluded images with blurred, large, and bright spots; defocus; foam; mucus; and food residue, which might have led to misjudgment. Therefore, we should appropriately add some pictures with specific features in the data preparation process to make our training data more comprehensive to ensure that it is closest to real clinical experience. Second, although this study was a multicenter validation study, participants were only from China. Future studies should incorporate endoscopic images from other countries to verify the performance of our model. Third, we did not enroll sufficient videos to test the effectiveness and ability of the GEADS because the information of a single still image is limited, which may lead to misdiagnosis and confusion. Therefore, we would like to add more videos to promote the clinical practice of the GEADS in future studies. In conclusion, by adopting endoscopic images from multiple centers, the GEADS can achieve high accuracy for diagnosis of upper GI diseases and lesion location.

ACKNOWLEDGMENTS

We express our heartfelt gratitude to the participants for their cooperation. We also thank F. Yang and Z. Chen for data processing and Y.-w. Xu for data download.

REFERENCES

1. Peery AF, Crockett SD, Murphy CC, et al. Burden and cost of gastrointestinal, liver, and pancreatic diseases in the United States: update 2018. Gastroenterology 2019;156:254-72.
2. Peery AF, Dellon ES, Lund J, et al. Burden of gastrointestinal disease in the United States: 2012 update. Gastroenterology 2012;143:1179-87.
3. Rentz AM, Battista C, Trudeau E, et al. Symptom and health-related quality-of-life measures for use in selected gastrointestinal disease studies: a review and synthesis of the literature. Pharmacoeconomics 2001;19:349-63.
4. Hamashima C. Systematic Review Group and Guideline Development Group for Gastric Cancer Screening Guidelines. Update version of the Japanese Guidelines for Gastric Cancer Screening. Jpn J Clin Oncol 2018;48:673-83.
5. Leung WK, Wu MS, Kakugawa Y, et al. Screening for gastric cancer in Asia: current evidence and practice. Lancet Oncol 2008;9:279-87.
6. Yoshida N, Doyama H, Yano T, et al. Early gastric cancer detection in high-risk patients: a multicentre randomised controlled trial on the effect of second-generation narrow band imaging. Gut 2021;70:67-75.
7. Wu L, Zhou W, Wan X, et al. A deep neural network improves endoscopic detection of early gastric cancer without blind spots. Endoscopy 2019;51:522-31.
8. Menon S, Trudgill N. How commonly is upper gastrointestinal cancer missed at endoscopy? A meta-analysis. Endosc Int Open 2014;2:E46-50.
9. Horiuchi Y, Tokai Y, Yamamoto N, et al. Additive effect of magnifying endoscopy with narrow-band imaging for diagnosing mixed-type early gastric cancers. Dig Dis Sci 2020;65:591-9.
10. Flynn AD, Valentine JF. Chromoendoscopy for dysplasia surveillance in inflammatory bowel disease. Inflamm Bowel Dis 2018;24:1440-52.
11. Yao K, Uedo N, Muto M, et al. Development of an e-learning system for teaching endoscopists how to diagnose early gastric cancer: basic principles for improving early detection. Gastric Cancer 2017;20:28-38.
12. Horiuchi Y, Hirasawa T, Ishizuka N, et al. Performance of a computer-aided diagnosis system in diagnosing early gastric cancer using magnifying endoscopy videos with narrow-band imaging (with videos). Gastrointest Endosc 2020;92:856-65.
13. Hewett DG, Kaltenbach T, Sano Y, et al. Validation of a simple classification system for endoscopic diagnosis of small colorectal polyps using narrow-band imaging. Gastroenterology 2012;143:599-607.
14. Yoon HJ, Kim S, Kim JH, et al. A lesion-based convolutional neural network improves endoscopic detection and depth prediction of early gastric cancer. J Clin Med 2019;8:1310-9.
15. Hirasawa T, Aoyama K, Tanimoto T, et al. Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images. Gastric Cancer 2018;21:653-60.
16. Zhang Y, Li F, Yuan F, et al. Diagnosing chronic atrophic gastritis by gastroscopy using artificial intelligence. Dig Liver Dis 2020;52:566-72.
17. Guimaraes P, Keller A, Fehlmann T, et al. Deep-learning based detection of gastric precancerous conditions. Gut 2020;69:4-6.
18. Shichijo S, Nomura S, Aoyama K, et al. Application of convolutional neural networks in the diagnosis of *Helicobacter pylori* infection based on endoscopic images. EBioMedicine 2017;25:106-11.
19. Nakashima H, Kawahira H, Kawachi H, et al. Endoscopic three-categorical diagnosis of *Helicobacter pylori* infection using linked color imaging and deep learning: a single-center prospective study (with video). Gastric Cancer 2020;23:1033-40.
20. Zheng W, Zhang X, Kim JJ, et al. High accuracy of convolutional neural network for evaluation of *Helicobacter pylori* infection based on endoscopic images: preliminary experience. Clin Transl Gastroenterol 2019;10:e00109.
21. Luo H, Xu G, Li C, et al. Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: a multicentre, case-control, diagnostic study. Lancet Oncol 2019;20:1645-54.
22. Mori Y, Kudo SE, Misawa M, et al. Real-time use of artificial intelligence in identification of diminutive polyps during colonoscopy: a prospective study. Ann Intern Med 2018;169:357-66.
23. Takenaka K, Ohtsuka K, Fujii T, et al. Development and validation of a deep neural network for accurate evaluation of endoscopic images from patients with ulcerative colitis. Gastroenterology 2020;158:2150-7.
24. Cho BJ, Bang CS, Lee JJ, et al. Prediction of submucosal invasion for gastric neoplasms in endoscopic images using deep-learning. J Clin Med 2020;9:1858-71.
25. Horiuchi Y, Aoyama K, Tokai Y, et al. Convolutional neural network for differentiating gastric cancer from gastritis using magnified endoscopy with narrow band imaging. Dig Dis Sci 2020;65:1355-63.
26. Shichijo S, Endo Y, Aoyama K, et al. Application of convolutional neural networks for evaluating *Helicobacter pylori* infection status on the basis of endoscopic images. Scand J Gastroenterol 2019;54:158-63.
27. 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.
28. 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.
29. de Groof AJ, Struyvenberg MR, van der Putten J, et al. Deep-learning system detects neoplasia in patients with Barrett's esophagus with higher accuracy than endoscopists in a multistep training and validation study with benchmarking. Gastroenterology 2020;158:915-29.
30. Tang D, Wang L, Ling T, et al. Development and validation of a real-time artificial intelligence-assisted system for detecting early gastric cancer: a multicentre retrospective diagnostic study. EBioMedicine 2020;62:103146.
31. Li L, Chen Y, Shen Z, et al. Convolutional neural network for the diagnosis of early gastric cancer based on magnifying narrow band imaging. Gastric Cancer 2020;23:126-32.
32. He X, Wu L, Dong Z, et al. Real-time use of artificial intelligence for diagnosing early gastric cancer by magnifying image-enhanced

endoscopy: a multicenter diagnostic study (with videos). Gastrointest Endosc 2022;95:671-8.

Abbreviations: AI, artificial intelligence; AUC, area under the curve; CI, confidence interval; GEADS, GI endoscopy artificial intelligence–assisted diagnosis system.

DISCLOSURE: *All authors disclosed no financial relationships. Research support for this study was provided by the National Key Research and Development Program of China (2018AAA0100201), National Natural Science Foundation of China (81902861), Innovative Scientific Research Project of Medical Youth in Sichuan Province (Q20073), Dazhou-Sichuan University Intelligent Medical Laboratory in Dazhou (2020CDDZ-02), The Key Projects fund of Science & Technology Department of Sichuan Province (22ZDYF1942, 227DYE1766), and Dazhou Science and Technology Bureau (21ZDYF0029).*

*Mr H. Yang, Ms Y. Wu, Mr B. Yang, and Ms M. Wu contributed equally to this article.

0016-5107/$36.00
https://doi.org/10.1016/j.gie.2022.06.011

Received January 30, 2022. Accepted June 11, 2022.

Current affiliations: Department of Clinical Research Center (1), Digestive Endoscopy Center (3), Department of Cardiac &Vascular Surgery (10), Dazhou Central Hospital, Dazhou, Sichuan, China; Center of Intelligent Medicine, Computer Science (2), Huaxi MR Research Center (HMRRC), Department of Radiology, West China Hospital (4), Precision Medicine Center, West China Hospital (5), Sichuan University, Chengdu, Sichuan, China; College of Aulin, Northeast Forestry University, Harbin, Heilongjiang, China (6), College of Basic Medical Sciences, North Sichuan Medical College, Nanchong, Sichuan, China (7), Department of Surgery, Shantou University Medical College, Shantou, Guangdong, China (8), Department of Sichuan, Academy of Medical Sciences, Sichuan Provincial People's Hospital, Chengdu, Sichuan, China (9), Department of Spinal Surgery, Sichuan Province Orthopedic Hospital, Chengdu, Sichuan, China (11), Department of Hepatobiliary Surgery, State Key Laboratory of Molecular Oncology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing, China (12).

Correspondence to: Hong Zhao at zhaohong@cicams.ac.cn.or

Reprint requests: Fanxin Zeng, PhD, Department of Clinical Research Center, Dazhou Central Hospital, No. 56 Nanyuemiao St, Tongchuan District, Dazhou 635000, China.

SUPPLEMENTARY METHODS

Development of the GI endoscopy artificial intelligence–assisted diagnosis system algorithm

In the gastroscopic image, the size of lesion is usually not obvious, especially for polyps, which requires the deep neural network models to have strong multiscale feature expression ability. For this reason, StoNet was proposed based on the concept of DenseNet121 and Feature Pyramid Network (FPN) to reduce the ignored features by deep architecture.[1,2] In StoNet, multiscale features can be reused, and high-dimension features are merged with low-dimension features to enrich the semantics of all scale features. It is proven to be benefit for detecting diseases in medical images.

StoNet consists of 3 stages. The first stage is designed for feature extraction, the second stage is used for feature fusion, and the last stage is feature classification. At the stage of feature extraction, DensNet121 is used as a backbone to extract features with different scales. The reason for choosing DenseNet121 as a backbone is that convolutional layers in dense blocks of DenseNet are densely connected, which means every convolutional layer could receive all output features of its preceding convolutional layers. This structure makes DenseNet have a powerful feature expression ability with no need for increasing the number of parameters. DenseNet121 has 4 dense blocks with 4 kinds of output feature sizes.[3]

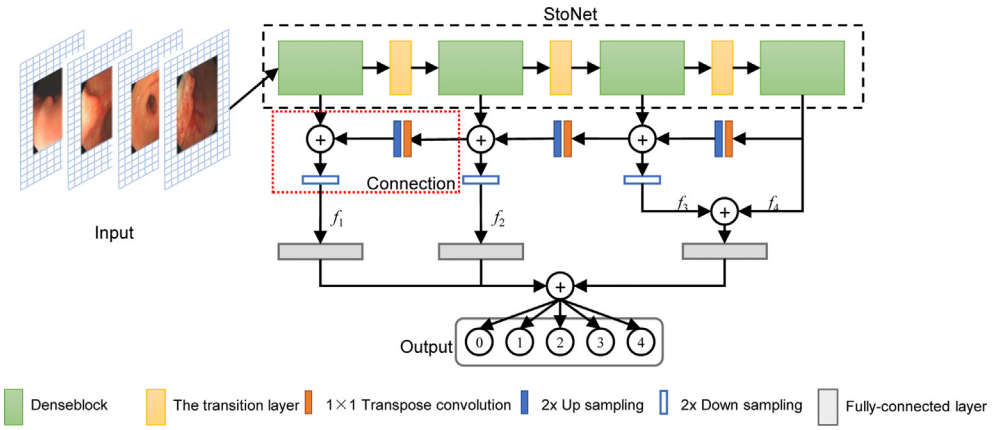
The second stage is used for feature fusion. In convolutional neural networks, the scale of high-dimensional features is small, which contains rich semantics, but the resolution is low. Meanwhile, low-dimensional features have a larger scale and higher resolution but less semantic features.[4] At the stage of feature fusion, to enrich the semantics of all scale features, a feature fusion connection was proposed to combine adjacent high-dimension features with low-dimension features in StoNet. In feature connection, a transposed convolution layer with a 1×1 kernel size is used first to reduce the dimension of high-dimension features, and then upsampling by factor of 2 is used subsequently to magnify the scale of high-dimension features so that the high-dimension features have the same scale as low-dimension features. The upsampling method used in this stage is bilinear interpolation. Because of the smoothing effect of bilinear interpolation, this upsampling method will cause contour detail degradation to a certain extent when the high-dimensional features are enlarged. The transposed convolution layer with a 1×1 kernel size used before upsampling could also make the network learn the appropriate variable feature transformation automatically to reduce the detail degradation caused by bilinear interpolation. Then high-dimension features and low-dimension features are added.[5] Finally, through dowsampling by a factor of 2, the high-dimension features and low-dimension features are fused completely. The dowsampling method used in this stage is bilinear interpolation. If 2 top-most features have the same scale, we add them up again to make the features contain more information.

At the stage of feature classification, 3 fully convolution layers are used as classifiers to classify the features with different scales into 5 stomach diseases separately, in which 0 means normal, 1 means cancer, 2 means erosion, 3 means polyp, and 4 means ulcer. The ultimate outcome of StoNet is determined by the outputs of these 3 full convolution layers in a weighted voting way. These 3 classifiers output the probabilities of the 5 stomach diseases separately, and then the probabilities of the same disease were added up, and the disease with the highest final probability was chosen as the output of StoNet.

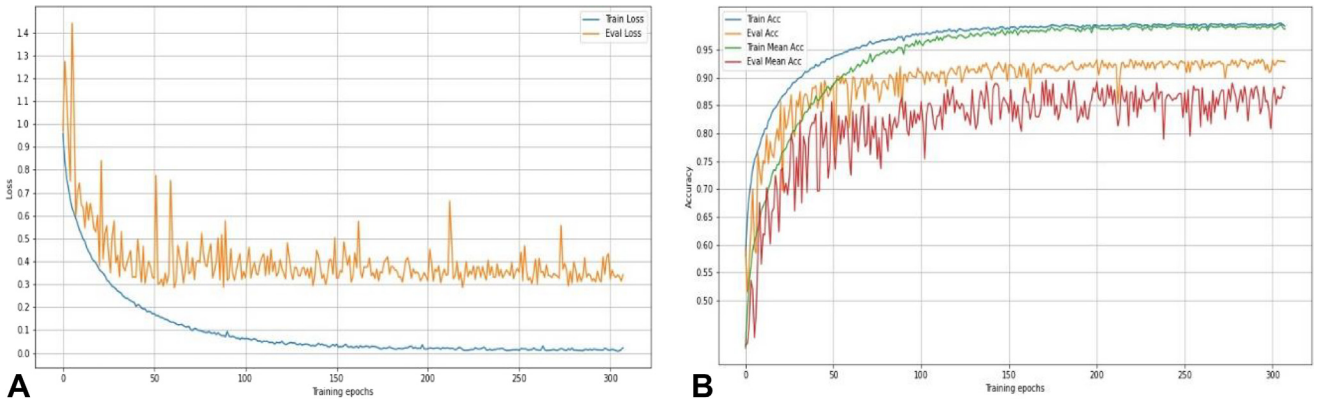
At the training of StoNet, cross-entropy was chosen as the loss function, batch size was set to 100, Stochastic Gradient Descent (SGD) was used to optimize the parameters of StoNet, and the learning rate was .001. We adopted random rotation, random vertical flipping, and random horizontal flipping as data augmentations to enrich training data. The training of StoNet has 2 periods. The first period is used to train a DenseNet121 as a pretrained model of StoNet, and then the trained DenseNet121 is used as the backbone of StoNet by transfer learning. The training epochs in this period are 300. The second period is used to train the StoNet integrally, and the training epochs in this period are 50.

REFERENCES

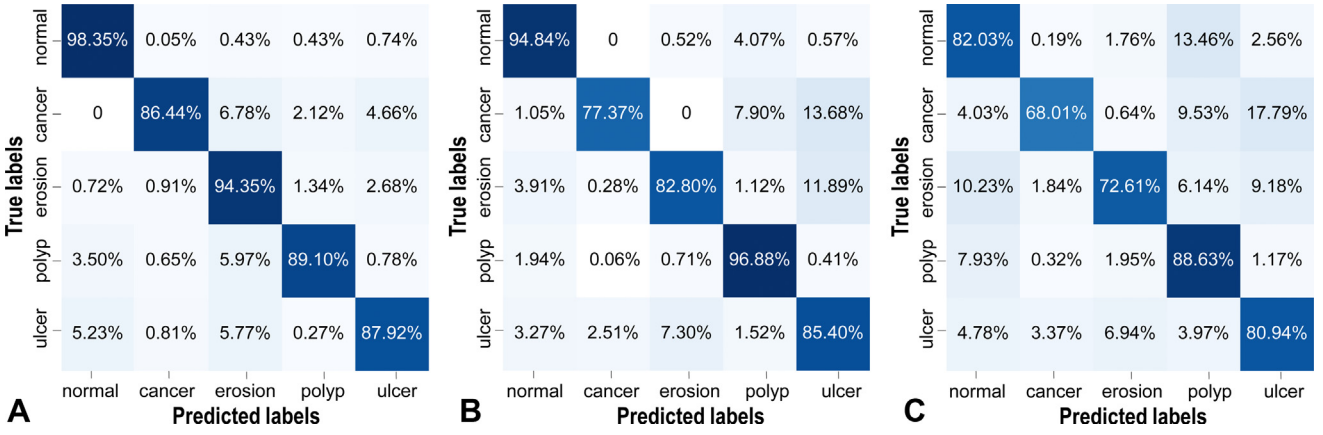
1. Huang G, Liu Z, Laurens V, et al. Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, USA. p. 2261-9.
2. Li H, Ouyang W, Wang X. Multi-bias non-linear activation in deep neural networks. arXiv 2016:1604.00676.
3. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv 2014:1409.1556.
4. Xuan Y, Jing P, Rister BB, et al. a systematic approach to blocking convolutional neural networks. arXiv 2016:1606.04209.
5. Jégou S, Drozdzal M, Vazquez D, et al. The one hundred layers tiramisu: fully convolutional DenseNets for semantic segmentation. IEEE 2017;1:1175-83.



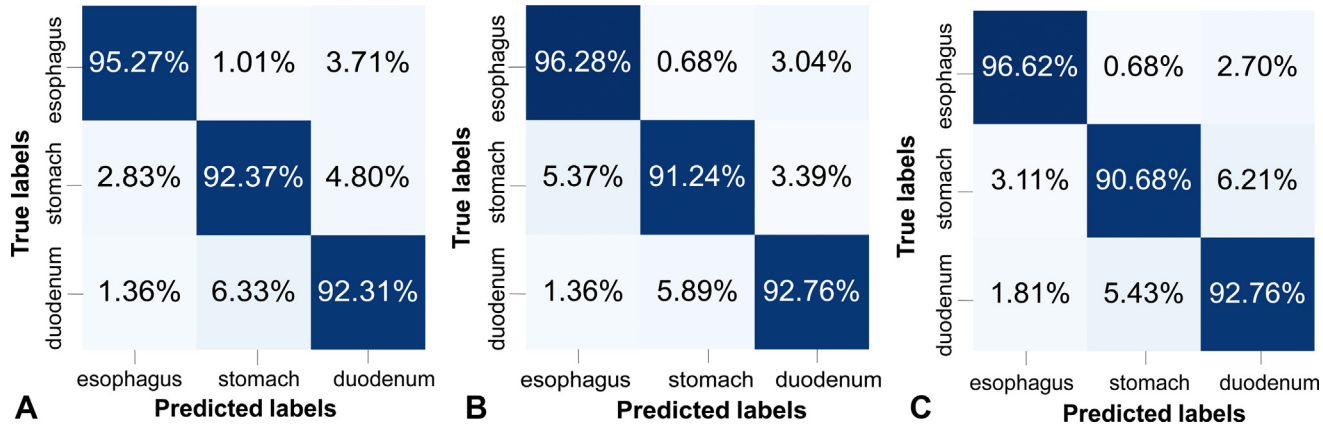
Supplementary Figure 1. Pattern of the StoNet.



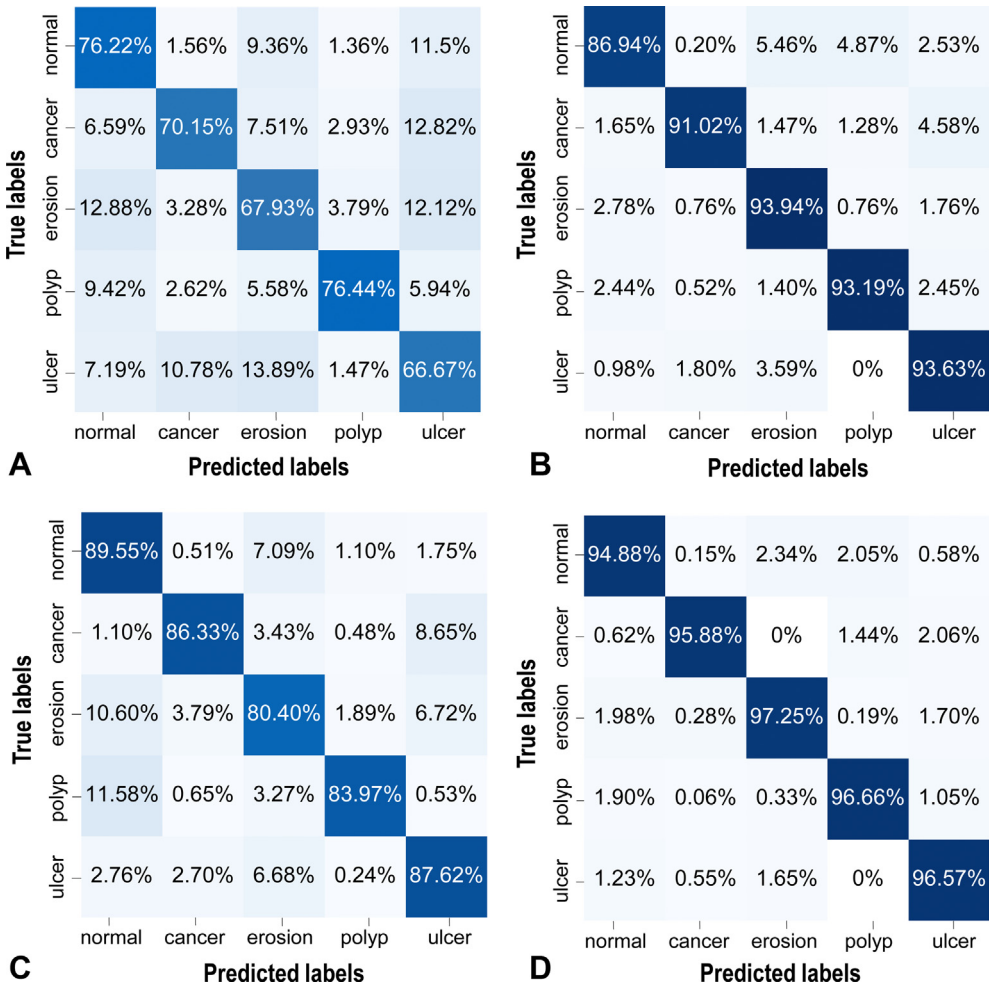
Supplementary Figure 2. The training process of the StoNet model. **A,** The loss of cross-entropy in the StoNet model. **B,** The improvement of accuracy in the StoNet model.



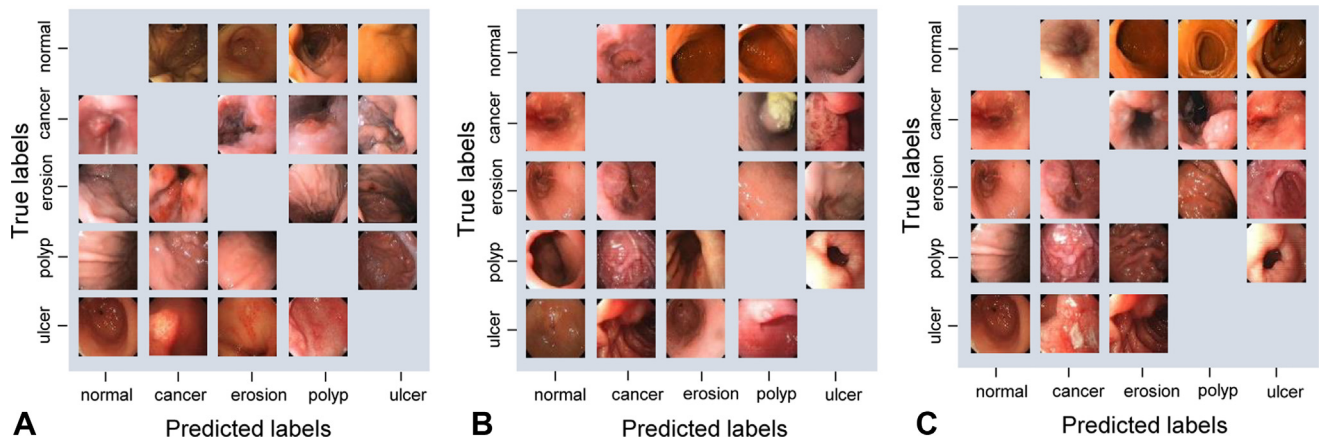
Supplementary Figure 3. Diagnostic performance of the StoNet model in 5 classification tasks. **A,** Confusion matrix of the StoNet model in the training dataset. **B,** Confusion matrix of the StoNet model in the validation dataset. **C,** Confusion matrix of the StoNet model in the test dataset.



Supplementary Figure 4. Diagnostic performance of anatomic locations for different models in the test dataset. **A,** Confusion matrix of the DenseNet121 model. **B,** Confusion matrix of the ResNet50 model. **C,** Confusion matrix of the VGG-11 model.



Supplementary Figure 5. Comparison of diagnostic performance between the artificial intelligence system and endoscopists. **A,** Novices. **B,** GI endoscopy artificial intelligence–assisted diagnosis system (GEADS) and novices. **C,** Experts with >5 years of experience. **D,** GEADS and experts with >5 years of experience.



Supplementary Figure 6. Performance of GI endoscopy artificial intelligence–assisted diagnosis system (GEADS) compared with endoscopists in identifying upper GI diseases in testing images from the comparison dataset. **A,** Examples of endoscopic images with discordant assessments by the endoscopist. **B,** Examples of endoscopic images with discordant assessments by the GEADS model. **C,** Examples of endoscopic images with discordant assessments by the endoscopist and the GEADS model.

SUPPLEMENTARY TABLE 1. Datasets for developing anatomic localization models

Location	Training	Validation	Test
Esophagus	890 (34.0)	296 (34.0)	296 (34.0)
Stomach	1061 (40.6)	354 (40.6)	354 (40.6)
Duodenum	663 (25.4)	221 (25.4)	221 (25.4)

Values are n (%).

SUPPLEMENTARY TABLE 2. Information of upper GI cancer patients in Dazhou Central Hospital and external validation datasets

Characteristics	Dazhou Central Hospital validation	External validation
Location		
Esophagus	247 (53.6)	745 (52.0)
Gastric	214 (46.4)	689 (48.0)
Pathology		
Low-grade intraepithelial neoplasia	358 (77.7)	1082 (75.5)
High-grade intraepithelial neoplasia	61 (13.2)	190 (13.2)
Submucosal invasion by carcinoma	42 (9.1)	162 (11.3)

Values are n (%).

SUPPLEMENTARY TABLE 3. Performance of GEADS in different validation datasets

	Dazhou Central Hospital validation			External validation					
	Validation	Test	Prospective	Xuanhan People's Hospital	Quxian People's Hospital	Kaijiang People's Hospital	Tongchuan People's Hospital	Dachuan People's Hospital	Dazhu People's Hospital
Accuracy	.918 (.914-.922)	.916 (.910-.920)	.925 (.922-.929)	.938 (.931-.945)	.841 (.834-.848)	.892 (.872-.912)	.949 (.935-.963)	.889 (.874-.904)	.895 (.884-.906)
F1 score	.884 (.879-.889)	.883 (.878-.888)	.845 (.840-.849)	.916 (.908-.924)	.828 (.821-.835)	.842 (.819-.865)	.886 (.866-.906)	.856 (.839-.872)	.890 (.879-.901)
Recall	.873 (.868-.878)	.875 (.870-.880)	.910 (.906-.913)	.912 (.904-.920)	.829 (.822-.836)	.818 (.793-.842)	.854 (.831-.876)	.851 (.834-.868)	.889 (.879-.901)
Precision	.890 (.885-.895)	.893 (.888-.898)	.807 (.801-.812)	.922 (.914-.930)	.828 (.821-.835)	.873 (.852-.894)	.926 (.909-.942)	.861 (.844-.877)	.891 (.880-.903)

Values are mean (95% confidence interval). *GEADS*, GI endoscopy artificial intelligence–assisted diagnosis system.

SUPPLEMENTARY TABLE 4. Summary of the performance by different convolutional neural network models in discriminating normal and upper GI diseases

Model	Items	Accuracy	F1 score	Recall	Precision
StoNet	Normal	.965 (.960-.970)	.948 (.942-.954)	.949 (.943-.954)	.947 (.941-.953)
	Cancer	.987 (.984-.990)	.810 (.799-.821)	.774 (.762-.785)	.850 (.840-.859)
	Erosion	.960 (.955-.965)	.849 (.839-.858)	.828 (.818-.838)	.871 (.862-.880)
	Polyp	.969 (.965-.974)	.953 (.948-.959)	.969 (.964-.974)	.939 (.932-.945)
	Ulcer	.950 (.944-.956)	.857 (.847-.866)	.854 (.845-.864)	.860 (.850-.869)
VGG-11	Normal	.913 (.906-.921)	.856 (.847-.866)	.780 (.768-.791)	.950 (.944-.956)
	Cancer	.989 (.986-.992)	.839 (.829-.849)	.821 (.811-.831)	.857 (.848-.867)
	Erosion	.947 (.941-.953)	.801 (.791-.812)	.785 (.774-.796)	.819 (.809-.829)
	Polyp	.941 (.935-.948)	.912 (.904-.920)	.944 (.939-.951)	.881 (.873-.890)
	Ulcer	.913 (.906-.921)	.779 (.768-.790)	.876 (.867-.885)	.701 (.689-.713)
ResNet50	Normal	.948 (.942-.954)	.924 (.916-.931)	.943 (.937-.950)	.904 (.897-.912)
	Cancer	.986 (.982-.989)	.808 (.797-.819)	.842 (.832-.852)	.777 (.766-.788)
	Erosion	.949 (.943-.955)	.800 (.789-.811)	.748 (.737-.760)	.860 (.851-.870)
	Polyp	.961 (.956-.966)	.940 (.933-.946)	.943 (.937-.949)	.936 (.930-.943)
	Ulcer	.942 (.935-.948)	.831 (.821-.841)	.827 (.817-.837)	.836 (.826-.846)
DenseNet121	Normal	.956 (.951-.962)	.933 (.927-.940)	.921 (.914-.928)	.946 (.940-.952)
	Cancer	.985 (.982-.989)	.785 (.773-.796)	.747 (.736-.759)	.826 (.815-.836)
	Erosion	.955 (.950-.961)	.839 (.830-.849)	.859 (.849-.868)	.821 (.811-.831)
	Polyp	.963 (.958-.968)	.943 (.937-.949)	.957 (.952-.963)	.929 (.922-.936)
	Ulcer	.943 (.936-.949)	.834 (.824-.844)	.825 (.814-.835)	.843 (.833-.853)

Values are mean (95% confidence interval).

SUPPLEMENTARY TABLE 5. Diagnostic performance of anatomic location by different models

Model	Items	Accuracy	F1 score	Recall	Precision
VGG-11	Esophagus	.971 (.960-.982)	.958 (.945-.971)	.966 (.954-.978)	.950 (.936-.965)
	Stomach	.946 (.931-.961)	.932 (.915-.949)	.907 (.888-.926)	.958 (.945-.972)
	Duodenum	.947 (.932-.962)	.899 (.879-.919)	.928(.910-.945)	.872 (.850-.895)
ResNet50	Esophagus	.962 (.949-.975)	.945 (.930-.960)	.963 (.950-.975)	.928 (.911-.946)
	Stomach	.947 (.932-.962)	.934 (.917-.950)	.912 (.894-.931)	.956 (.942-.969)
	Duodenum	.958 (.944-.971)	.917 (.899-.936)	.928 (.910-.945)	.907 (.888-.926)
DenseNet121	Esophagus	.969 (.958-.981)	.954 (.940-.968)	.953 (.939-.967)	.956 (.942-.970)
	Stomach	.950 (.935-.964)	.937 (.920-.953)	.924 (.906-.941)	.951 (.936-.965)
	Duodenum	.948 (.934-.963)	.901 (.881-.921)	.923 (.905-.941)	.879 (.858-.901)

Values are mean (95% confidence interval).

SUPPLEMENTARY TABLE 6. GEADS model and endoscopists in discriminating normal, cancer, erosion, polyps, and ulcers

	Items	Accuracy	F1 score	Recall	Precision	False-positive rate (%)	False-negative rate (%)	*P* value
GEADS	AI model	.893 (.873-.914)	.893 (.872-.913)	.891 (.870-.911)	.901 (.881-.920)	2.7	11.0	
Novices	Novice 1	.696 (.665-.726)	.693 (.662-.723)	.706 (.676-.736)	.706 (.676-.736)	7.5	29.4	.034*
	Novice 2	.749 (.720-.778)	.726 (.696-.755)	.729 (.700-.759)	.773 (.745-.801)	6.3	27.1	
	Novice 3	.702 (.672-.733)	.704 (.674-.734)	.709 (.679-.739)	.763 (.735-.791)	7.4	29.1	
Experts	Expert 1	.885 (.864-.906)	.883 (.862-.904)	.886 (.865-.907)	.889 (.868-.909)	2.8	11.4	.557*
	Expert 2	.878 (.857-.900)	.875 (.853-.897)	.878 (.857-.900)	.880 (.858-.901)	3.0	12.2	
	Expert 3	.916 (.898-.934)	.914 (.896-.933)	.916 (.898-.935)	.916 (.897-.934)	2.1	8.4	
	Expert 4	.856 (.833-.879)	.853 (.829-.876)	.852 (.829-.875)	.868 (.846-.890)	3.6	14.8	
	Expert 5	.896 (.875-.916)	.892 (.871-.912)	.897 (.877-.917)	.894 (.873-.914)	2.5	10.3	
	Expert 6	.868 (.846-.890)	.862 (.839-.885)	.862 (.840-.885)	.871 (.849-.893)	3.3	13.8	
	Expert 7	.749 (.720-.778)	.744 (.715-.773)	.750 (.721-.778)	.755 (.726-.783)	6.2	25.0	
	Expert 8	.821 (.795-.846)	.807 (.781-.833)	.804 (.778-.831)	.841 (.817-.865)	4.5	19.6	
Novices with GEADS	Novice 1	.919 (.901-.937)	.918 (.900-.937)	.919 (.901-.937)	.920 (.902-.938)	2.0	8.1	<.001†
	Novice 2	.905 (.885-.924)	.902 (.882-.922)	.905 (.886-.924)	.902 (.882-.921)	2.4	9.5	
	Novice 3	.928 (.911-.945)	.927 (.909-.944)	.928 (.911-.945)	.927 (.910-.944)	1.8	7.2	
Experts with GEADS	Expert 1	.957 (.943-.970)	.956 (.943-.970)	.957 (.944-.970)	.956 (.942-.969)	1.1	4.3	<.001†
	Expert 2	.950 (.936-.964)	.949 (.935-.964)	.949 (.934-.963)	.950 (.934-.965)	1.3	5.1	
	Expert 3	.966 (.954-.978)	.966 (.954-.978)	.967 (.955-.978)	.966 (.954-.978)	.9	3.4	
	Expert 4	.960 (.947-.973)	.960 (.947-.973)	.959 (.946-.973)	.960 (.947-.973)	1.0	4.1	
	Expert 5	.963 (.950-.975)	.962 (.949-.974)	.963 (.950-.975)	.961 (.948-.974)	.9	3.7	
	Expert 6	.969 (.958-.981)	.969 (.957-.980)	.970 (.959-.982)	.968 (.956-.980)	.8	3.0	
	Expert 7	.968 (.957-.980)	.967 (.956-.979)	.969 (.957-.980)	.966 (.954-.978)	.8	3.1	
	Expert 8	.965 (.953-.977)	.964 (.952-.977)	.966 (.955-.978)	.963 (.950-.975)	.9	3.3	

Values are mean (95% confidence interval) unless otherwise defined.
GEADS, GI endoscopy artificial intelligence–assisted diagnosis system.
*Comparing accuracy differences between GEADS and endoscopists.
†Comparing accuracy differences between endoscopists and endoscopists with GEADS.