THERAPEUTIC ADVANCES in
Gastroenterology

Original Research

An explainable artificial intelligence system for diagnosing *Helicobacter Pylori* infection under endoscopy: a case–control study

Mengjiao Zhang*, Jie Pan*, Jiejun Lin, Ming Xu, Lihui Zhang, Renduo Shang, Liwen Yao, Yanxia Li, Wei Zhou, Yunchao Deng, Zehua Dong, Yijie Zhu, Xiao Tao, Lianlian Wu and Honggang Yu

Ther Adv Gastroenterol

2023, Vol. 16: 1–14

DOI: 10.1177/17562848231155023

© The Author(s), 2023.
Article reuse guidelines:
sagepub.com/journals-permissions

Abstract

Background: Changes in gastric mucosa caused by *Helicobacter pylori* (*H. pylori*) infection affect the observation of early gastric cancer under endoscopy. Although previous researches reported that computer-aided diagnosis (CAD) systems have great potential in the diagnosis of *H. pylori* infection, their explainability remains a challenge.
Objective: We aim to develop an explainable artificial intelligence system for diagnosing *H. pylori* infection (EADHI) and giving diagnostic basis under endoscopy.
Design: A case–control study.
Methods: We retrospectively obtained 47,239 images from 1826 patients between 1 June 2020 and 31 July 2021 at Renmin Hospital of Wuhan University for the development of EADHI. EADHI was developed based on feature extraction combining ResNet-50 and long short-term memory networks. Nine endoscopic features were used for *H. pylori* infection. EADHI's performance was evaluated and compared to that of endoscopists. An external test was conducted in Wenzhou Central Hospital to evaluate its robustness. A gradient-boosting decision tree model was used to examine the contributions of different mucosal features for diagnosing *H. pylori* infection.
Results: The system extracted mucosal features for diagnosing *H. pylori* infection with an overall accuracy of 78.3% [95% confidence interval (CI): 76.2–80.3]. The accuracy of EADHI for diagnosing *H. pylori* infection (91.1%, 95% CI: 85.7–94.6) was significantly higher than that of endoscopists (by 15.5%, 95% CI: 9.7–21.3) in internal test. And it showed a good accuracy of 91.9% (95% CI: 85.6–95.7) in external test. Mucosal edema was the most important diagnostic feature for *H. pylori* positive, while regular arrangement of collecting venules was the most important *H. pylori* negative feature.
Conclusion: The EADHI discerns *H. pylori* gastritis with high accuracy and good explainability, which may improve the trust and acceptability of endoscopists on CADs.

Plain language summary

An explainable AI system for *Helicobacter pylori* with good diagnostic performance

Helicobacter pylori (*H. pylori*) is the main risk factor for gastric cancer (GC), and changes in gastric mucosa caused by *H. pylori* infection affect the observation of early GC under endoscopy. Therefore, it is necessary to identify *H. pylori* infection under endoscopy. Although previous research showed that computer-aided diagnosis (CAD) systems have great potential in *H. pylori* infection diagnosis, their generalization and explainability are

Correspondence to:
Honggang Yu
Department of Gastroenterology, Renmin Hospital of Wuhan University, No. 9 Zhangzhidong Road, Wuchang District, Wuhan, Hubei 430060, China

Hubei Key Laboratory of Digestive System Disease, Renmin Hospital of Wuhan University, Wuhan, Hubei, China

Hubei Provincial Clinical Research Center for Digestive Disease Minimally Invasive Incision, Renmin Hospital of Wuhan University, Wuhan, Hubei, China
yuhonggang@whu.edu.cn

Lianlian Wu
Department of Gastroenterology, Renmin Hospital of Wuhan University, No. 9 Zhangzhidong Road, Wuchang District, Wuhan, Hubei 430060, China

Hubei Key Laboratory of Digestive System Disease, Renmin Hospital of Wuhan University, Wuhan, Hubei, China

Hubei Provincial Clinical Research Center for Digestive Disease Minimally Invasive Incision, Renmin Hospital of Wuhan University, Wuhan, Hubei, China
wu_leanne@163.com

Mengjiao Zhang
Department of Gastroenterology, Renmin Hospital of Wuhan University, Wuhan, Hubei, China

Department of Gastroenterology, The Central Hospital of Wuhan, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

still a challenge. Herein, we constructed an explainable artificial intelligence system for diagnosing *H. pylori* infection (EADHI) using images by case. In this study, we integrated ResNet-50 and long short-term memory (LSTM) networks into the system. Among them, ResNet50 is used for feature extraction, LSTM is used to classify *H. pylori* infection status based on these features. Furthermore, we added the information of mucosal features in each case when training the system so that EADHI could identify and output which mucosal features are contained in a case. In our study, EADHI achieved good diagnostic performance with an accuracy of 91.1% [95% confidence interval (CI): 85.7–94.6], which was significantly higher than that of endoscopists (by 15.5%, 95% CI: 9.7–21.3%) in internal test. In addition, it showed a good diagnostic accuracy of 91.9% (95% CI: 85.6–95.7) in external tests. The EADHI discerns *H. pylori* gastritis with high accuracy and good explainability, which may improve the trust and acceptability of endoscopists on CADs. However, we only used data from a single center to develop EADHI, and it was not effective in identifying past *H. pylori* infection. Future, multicenter, prospective studies are needed to demonstrate the clinical applicability of CADs.

Keywords: explainable artificial intelligence, *Helicobacter pylori*, long short-term memory network, ResNet-50

Received: 18 August 2022; revised manuscript accepted: 18 January 2023.

Key Laboratory for Molecular Diagnosis of Hubei Province, The Central Hospital of Wuhan, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

Jie Pan
Jiejun Lin
Department of Gastroenterology, Wenzhou Central Hospital, Wenzhou, China

Ming Xu
Lihui Zhang
Renduo Shang
Liwen Yao
Yanxia Li
Wei Zhou
Yunchao Deng
Zehua Dong
Yijie Zhu
Xiao Tao
Department of Gastroenterology, Renmin Hospital of Wuhan University, Wuhan, Hubei, China

Hubei Key Laboratory of Digestive System Disease, Renmin Hospital of Wuhan University, Wuhan, Hubei, China

Hubei Provincial Clinical Research Center for Digestive Disease Minimally Invasive Incision, Renmin Hospital of Wuhan University, Wuhan, Hubei, China

*These authors contribute equally to this work.

Introduction

Gastric cancer (GC) is one of the most common malignancies, accounting for over 1,000,000 new cases and an estimated 783,000 deaths in 2018.[1] *Helicobacter pylori* (*H. pylori*) is the leading risk factor for GC, which induces atrophic gastritis and intestinal metaplasia, ultimately leading to the development of GC.[2–6] Furthermore, *H. pylori* eradication improves gastric mucosal atrophy and inhibits the development of intestinal metaplasia.[3,7] Therefore, early detection and eradication of *H. pylori* infection are essential to avoid the development of GC.

The most important tool for early GC (EGC) screening is white light endoscopy (WLE).[8] However, the risk stratification of EGC is related to endoscopic findings of *H. pylori* infection status. Atrophy, intestinal metaplasia, nodularity, etc., in *H. pylori*-positive patients are related to the risk of EGC, whereas atrophy plays a more important role in *H. pylori*-negative patients.[9] In addition, depressed macroscopic EGC lesions are more common in the infected cases than in the uninfected cases,[10] whereas flat elevated lesions are more common in uninfected cases than in the infected cases.[11] The mucosal hyperemia, edema, and redness caused by *H. pylori* infection make the surface and edges of EGC more difficult to observe.[12] Thus, recognizing *H. pylori* infection under endoscopy is critical for the diagnosis of EGC. *H. pylori* infection does not produce detectable specific lesions, making it difficult to diagnose using endoscopy.[13] In addition, the accuracy of endoscopists in diagnosing *H. pylori* gastritis based on previous experience was approximately 70% under WLE.[14,15] Fortunately, an increasing number of *H. pylori* infection-related mucosal features has been identified, allowing the diagnosis of *H. pylori* gastritis using WLE.[16–19] However, this approach requires advanced skills and knowledge.[20,21]

Recent studies have shown that artificial intelligence (AI) uses deep learning to achieve feature expression.[22] Furthermore, it plays a vital role in identifying upper gastrointestinal diseases, including esophageal cancer and EGC.[23,24] Researchers have made great efforts with the help of AI to diagnose *H. pylori* infection using WLE. Shichijo *et al.* collected 32,208 images from 1750 patients for convolutional neural network (CNN) model development and achieved an accuracy of 88.9%, which was higher than that of endoscopists.[25] Zheng *et al.*[26] developed the CNN model using 11,729 images with an accuracy of 81.4% for a single image per patient, and 93.8% for multiple images (8.3 ± 3.3) per patient, suggesting that CNN using multiple gastric images achieved

Table 1. Baseline characteristics according to dataset.

	Development dataset	Internal test dataset	External test dataset	*p* Value
No. of patients	1826	168	124	–
No. of images per endoscopy (mean ± SD)	25.87 ± 9.98	24.27 ± 8.01	41.18 ± 17.55	0.000
Age, years	46.55 ± 13.65	46.46 ± 12.69	48.73 ± 12.60	0.215
Sex, *n* (%)				0.591
Female	973 (53.29)	95 (56.55)	70 (56.45)	
Male	853 (46.71)	73 (43.45)	54 (43.55)	
H. pylori status, No. (%)				0.857
Positive	881 (48.25)	84 (50.00)	62 (50.00)	
Diagnosed by histological biopsy alone	91 (10.33)	10 (11.90)	0 (0)	
Negative	945 (51.75)	84 (50.00)	62 (50.00)	

SD, standard deviation.

higher diagnostic accuracy for the evaluation of *H. pylori* infection. However, previous diagnosis systems return a final decision result without explanation, making it difficult for endoscopists to learn from the models.[27] The poor explainability of these black-box models undermines the physicians' trust and puts patients at risk, severely limiting AI systems' clinical applications.[28,29] Therefore, improving the explainability of AI systems is necessary for their applications.

In this study, we developed an explainable AI system for diagnosing *H. pylori* infection (EADHI) based on feature extraction using ResNet-50 and long short-term memory network (LSTM). The performance of EADHI under WLE was evaluated in internal and external test sets and further compared with endoscopists of different levels. To the best of our knowledge, this is the first study to concretize abstract diagnostic theories through feature extraction, providing diagnostic results and a diagnostic basis for endoscopists.

Methods

Patients and esophagogastroduodenoscopy protocol

We retrospectively reviewed patients undergoing esophagogastroduodenoscopy (EGD) with gastric biopsies or *H. pylori* breath test at Renmin Hospital of Wuhan University (RHWU) between June 2020 and July 2021. We included 1826 patients (881 *H. pylori* positive and 945 *H. pylori* negative) for the development of EADHI. Table 1 shows the patient characteristics. Exclusion criteria include (1) patients with a history of GC, peptic ulcer, gastric surgery, or submucosal tumor and (2) patients who received *H. pylori* eradication or administered antibiotics within a month or proton pump inhibitor within 2 weeks of *H. pylori* breath test.

EGD was performed using a standard endoscope (GIF-HQ290, GIF-H260; Olympus, Tokyo, Japan; EG-L590ZW; Fujifilm, Tokyo, Japan) and the images were captured during high-definition, white-light examination of the antrum, angularis (retroflex), body (forward and retroflex), and fundus (retroflex). Gastric biopsies were performed in the antrum and body at the endoscopist's discretion.

Establishment of diagnostic feature for H. pylori infection with prior knowledge

Atrophy, intestinal metaplasia, nodularity, diffuse redness, spotty redness, mucosal swelling, enlarged folds, and sticky mucus are positive endoscopic findings for *H. pylori*. At the same time, regular arrangement of collecting venules (RAC), fundic gland polyp (FGP), red streak,

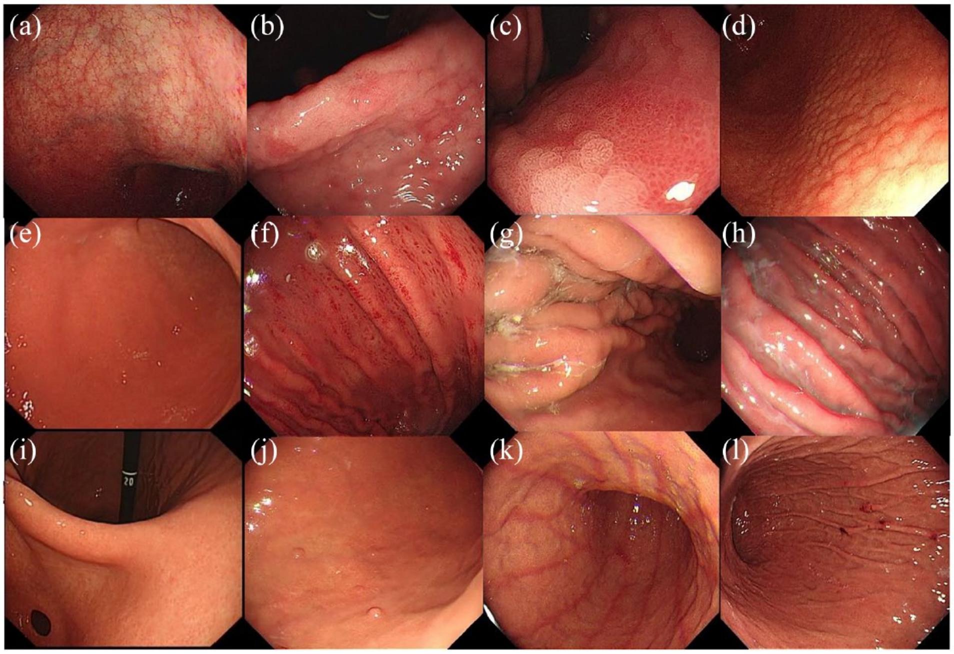

Figure 1. Evaluated endoscopic features: (a) atrophy, (b–c) intestinal metaplasia, (d) nodularity, (e–f) mucosal redness, (g–h) mucosal edema, (i) RAC, (j) FGP, (k) red streak, and (l) hematin.
FGP, fundic gland polyp; RAC, regular arrangement of collecting venules.

and hematin are predictive features for *H. pylori* negative.[16–19] These endoscopic features in the patients' images were evaluated by two experts, who had performed more than 5000 examinations. Before evaluating the images, the two experts were educated on the Kyoto classification of gastritis[16] using PowerPoint presentation. This pre-study training was conducted to avoid interobserver variance, making the endoscopist labels more objective and accurate. The two experts were blinded to the results of patients' *H. pylori* infection and were asked to evaluate the same images independently. We included those in which they both agreed and had distinct mucosal features. Mucosal swelling, enlarged folds, and sticky mucus frequently appear in the same images, and these features were collectively referred to as mucosal edema. Because spot redness and diffuse redness are similar and the background of spotty redness is mostly diffuse redness,[16] the two features were classified as mucosal redness. Supplemental Table S1 shows the results of the endoscopic findings in patients. A logistic regression model[30] was employed to evaluate the predictive capabilities of each endoscopic feature (Supplemental Table S2). According to the similarity between endoscopic features and their performance, nine features were finally included. Figure 1 shows the typical images.

Datasets and data preprocessing

Figure 2 shows the study's workflow. EGD images retrospectively obtained from RHWU were used for the development and internal test of EADHI. Meanwhile, to verify the robustness of the system, EGD images were collected from Wenzhou Central Hospital for the external test.

Two gastroenterology doctoral students who had mastered the basic operation and diagnosis knowledge of EGD removed images of duodenum and esophagus, poor quality, non-white light images, and those with active bleeding. Finally, 47,239 images from 1826 patients were used for the development of EADHI. Two experts independently screened 5780 images with distinct mucosal features for developing a model to enhance its ability to identify mucosal features. The images were randomly allocated to the training validation and testing dataset at a ratio of 8:1:1. Random projective transformations such as scaling, shearing, zooming, and horizontal flipping were applied to selected images. Furthermore, in AI model, training projective transformation is

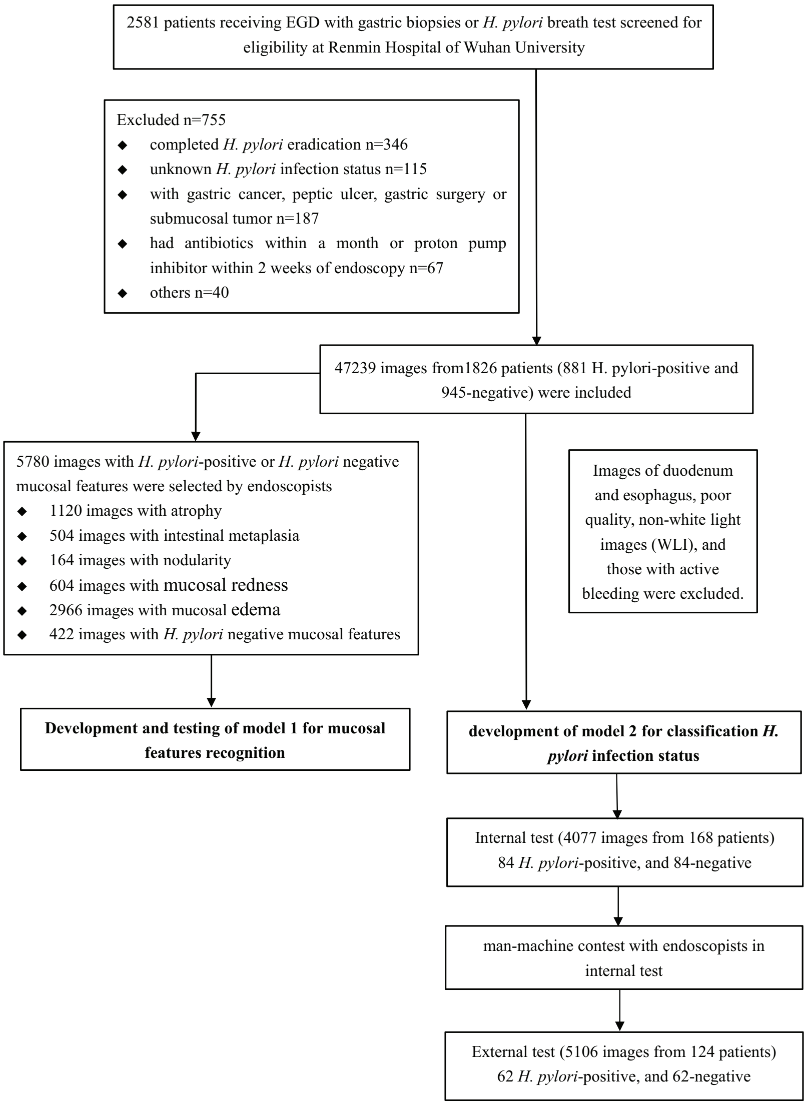
2581 patients receiving EGD with gastric biopsies or H. pylori breath test screened for eligibility at Renmin Hospital of Wuhan University
Excluded n=755
◆ completed H. pylori eradication n=346
◆ unknown H. pylori infection status n=115
◆ with gastric cancer, peptic ulcer, gastric surgery or submucosal tumor n=187
◆ had antibiotics within a month or proton pump inhibitor within 2 weeks of endoscopy n=67
◆ others n=40
47239 images from1826 patients (881 H. pylori-positive and 945-negative) were included
5780 images with H. pylori-positive or H. pylori negative mucosal features were selected by endoscopists
◆ 1120 images with atrophy
◆ 504 images with intestinal metaplasia
◆ 164 images with nodularity
◆ 604 images with mucosal redness
◆ 2966 images with mucosal edema
◆ 422 images with H. pylori negative mucosal features
Images of duodenum and esophagus, poor quality, non-white light images (WLI), and those with active bleeding were excluded.
Development and testing of model 1 for mucosal features recognition
development of model 2 for classification H. pylori infection status
Internal test (4077 images from 168 patients) 84 H. pylori-positive, and 84-negative
man-machine contest with endoscopists in internal test
External test (5106 images from 124 patients) 62 H. pylori-positive, and 62-negative

Figure 2. Workflow chart for the development and evaluation of EADHI.
EADHI, explainable artificial intelligence system for diagnosing *H. pylori* infection.

a commonly used data augmentation technique that can increase the robustness, stability, and generalization of the model.[24,31]

The prepared endoscopic images (about 26 images per case) of 881 *H. pylori*-positive patients and 945-negative patients (Table 1) were used for the development of model 2 to detect *H. pylori* infection based on identified mucosal features.

Internal testing dataset

A separate testing dataset was developed to evaluate the diagnostic accuracy of EADHI, and compare it with endoscopists. In August 2021, the EGD images (about 24 images per case) of 168 patients (84 *H. pylori* positive and 84 *H. pylori* negative) were included as the testing dataset and 71 patients were excluded using the exclusion criteria (Figure 3). Table 1 shows the patient demographics, and there was no overlap between the testing and the development datasets. In all, 10 endoscopists, trained on the Kyoto classification of gastritis[16] before assessing cases using PowerPoint presentation, of varying experience blinded to the patients' *H. pylori* infection status were independently asked whether a patient was *H. pylori* positive or *H. pylori* negative. Four of the 10 endoscopists were classified as follows: 'expert group', with EGDs > 5000. The other endoscopists

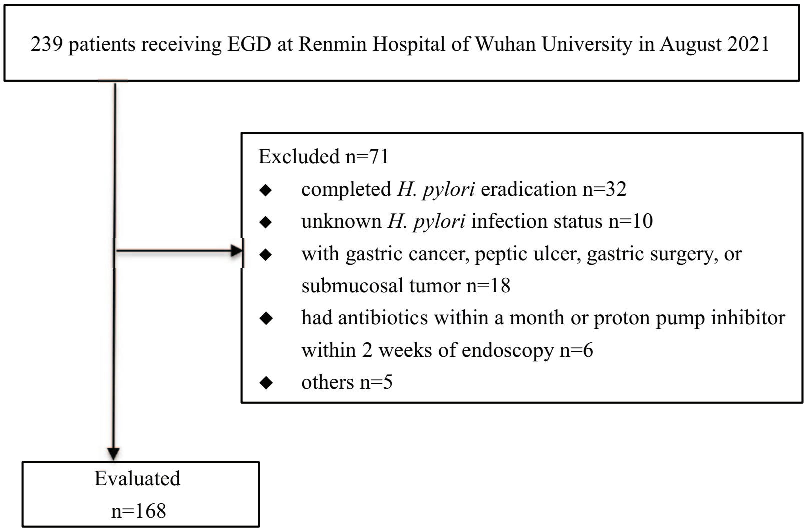
239 patients receiving EGD at Renmin Hospital of Wuhan University in August 2021
Excluded n=71
◆ completed *H. pylori* eradication n=32
◆ unknown *H. pylori* infection status n=10
◆ with gastric cancer, peptic ulcer, gastric surgery, or submucosal tumor n=18
◆ had antibiotics within a month or proton pump inhibitor within 2 weeks of endoscopy n=6
◆ others n=5
Evaluated n=168

Figure 3. Flow chart for the enrolled patients.

were further classified as the ‘relatively experienced group’, EGDs > 1000 (*n* = 3); and the ‘beginner group’, EGDs < 1000 (*n* = 3). All 10 endoscopists were not involved in the selection of the data.

External testing dataset

Images from the patients diagnosed with *H. pylori* positive or *H. pylori* negative (62 *H. pylori* positive and 62 *H. pylori* negative) were obtained from Wenzhou Central Hospital as an external testing dataset to evaluate the robustness of the EADHI (Table 1).

Gold standard of H. pylori *infection*

Patients who tested positive for *H. pylori* using histological examination of biopsy specimen or breath test were classified as *H. pylori* positive. However, only 9.8% (101/1027) of *H. pylori*-positive cases were diagnosed using histological examination of biopsy specimen alone (Table 1). Moreover, patients who tested negative on *H. pylori* breath test in the absence of eradiation treatment were classified as *H. pylori* negative.

Construction of the explainable AI system

The EADHI developed in the study contained two models, model 1 developed with ResNet-50 to extract mucosal features, and model 2 developed by combining model 1 with an LSTM network to detect *H. pylori* infection.

Model 1 was developed with ResNet-50 using six types of EGD images. Figure 4 shows a typical ResNet-50 architecture with these layers, including convolution layers, max-pooling Layers, 16 residual blocks, average pooling, and fully connected layers. High-level features obtained from the average pooling layer of trained ResNet-50 are fed into fully connected layers for classification. We further attempted to understand how model 1 recognized the input images by applying a Gradient-weighted Class Activation Map (Grad-CAM)[32] to determine which area of the images was most essential to the classification result. We developed heatmap images from the location map data. Supplemental Table S3 and Supplemental Figure S1 show the performance of model 1.

Model 2 was developed by combining ResNet-50 and LSTM networks, which were developed by the Keras deep learning framework (TensorFlow backend).[33] In model 2, ResNet-50, which incorporated all layers of model 1 except fully connected layers, was used for feature extraction, while LSTM was used to classify *H. pylori* infection status based on identified features. And the proposed model 2 developed in the study contains two phases, in phase one, ResNet-50 extracts all the mucosal features in a case from input case-based EGD images and generates feature vectors, whereas in phase two, LSTM receives the feature vectors to extract time information, which is then passed to the dense layer for classification (Figure 5).

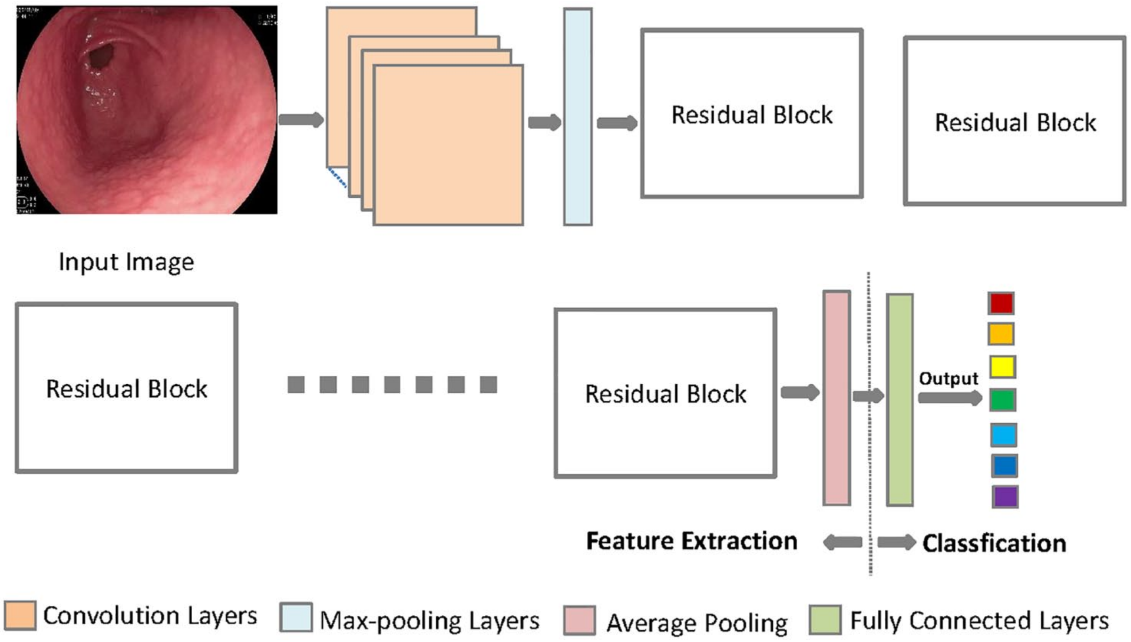

Figure 4. A typical architecture of model 1.

To make the diagnosis of the system more specific, we added the information about mucosal features when training model 2, and trained each feature separately. Supplemental Table S4 shows their performance in the validation dataset. As shown in Figure 6, EADHI could identify and output which mucosal features were present in a case, and then output the final result based on the identified mucosal features and their weights calculated using a gradient-boosting decision tree.[34] The sensitivity was plotted against the false-positive rate (i.e. 1-specificity) for all thresholds in the range [0, 1]. In addition, the receiver operating characteristic (ROC) curves were obtained.

Sample size

Sample size calculations were performed according to Exact Clopper–Pearson, with assumptions that EADHI has a sensitivity of 90% and a specificity of 90% for detecting *H. pylori* infection. A prevalence of 44% was estimated, based on recent demographic meta-analysis for China.[35] A sample size of 79 patients was calculated using a two-sided 95% confidence interval (CI) with a width of 0.1.

Ethics

We de-identified all patient information before data analysis to keep patients anonymous. Patient details were not accessible to any of the endoscopists involved in the study. The reporting of this study adheres to the Strengthening the Reporting of Observational Studies in Epidemiology statement.[36]

Statistical analysis

Demographic data were expressed as mean with standard deviation (SD). The performance of EADHI was evaluated using the following metrics: accuracy, specificity, sensitivity, positive predictive value, negative predictive value, and area under the curve (AUC) using ROC. Optimal cutoff values to obtain the highest AUC were calculated using the Youden index. The performance of the EADHI and endoscopists were compared using a two-tailed unpaired heteroscedastic Student’s t test. Two-sided $p < 0.05$ was considered statistically significant.

Results

Performance of EADHI for mucosal features

Table 2 shows that EADHI identified various mucosal features with an overall accuracy of 78.3% (95% CI: 76.2–80.3%), and the accuracy for atrophy, intestinal metaplasia, nodularity, mucosal redness, mucosal edema, RAC,

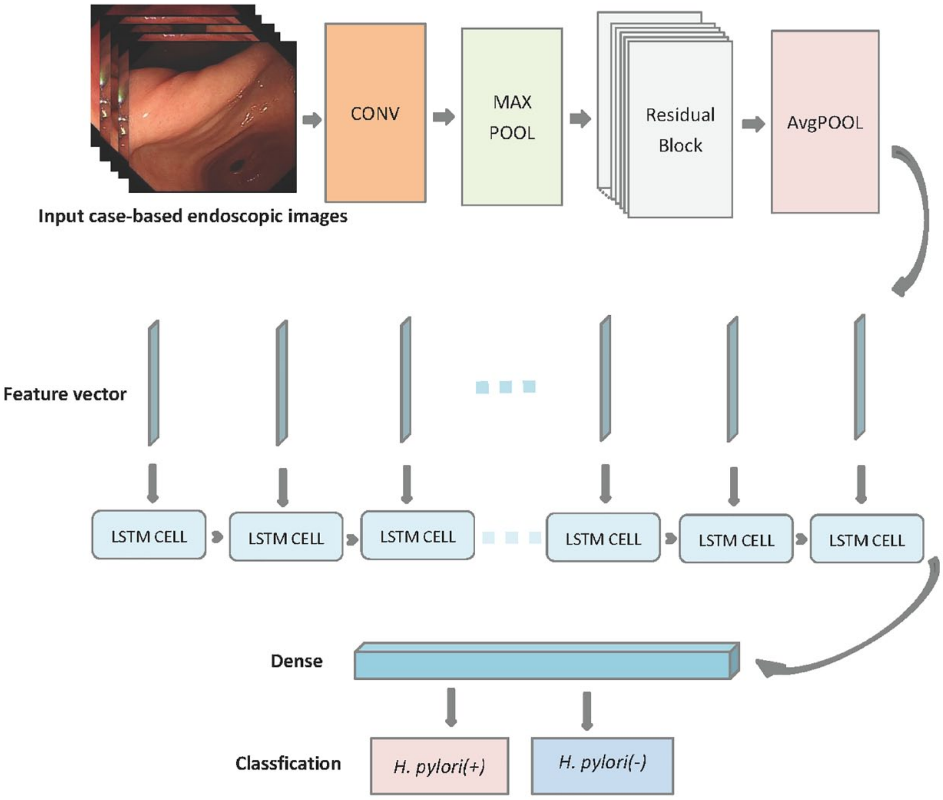

Figure 5. A typical architecture of model 2.

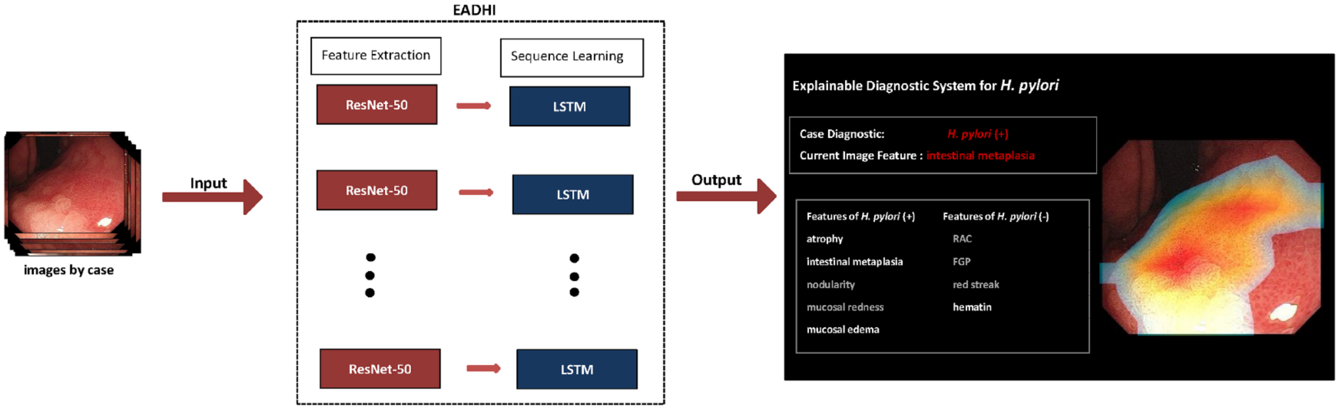

Figure 6. A schematic illustration of how EADHI diagnoses *H. pylori* infection in images by case. The ResNet-50 could identify and output mucosal features in a case, and the LSTM could diagnose *H. pylori* infection status based on identified mucosal features.
EADHI, explainable artificial intelligence system for diagnosing *H. pylori* infection; LSTM, long short-term memory.

FGP, red streak, and hematin were 78.6% (95% CI: 71.7–84.1%), 81.6% (95% CI: 75.0–86.7%), 91.7% (95% CI: 86.4–95.1%), 83.3% (95% CI: 76.9–88.3%), 85.7% (95% CI: 79.6–90.3%), 82.7% (95% CI: 76.3–87.8%), 69.6% (95% CI: 62.3–76.1%), 72.6% (95% CI: 65.4–78.8%), and 58.9% (95% CI: 51.4–66.1%). The AUC of EADHI for each mucosal feature ranged from 0.66 to 0.95 (Figure 7(a)). And the corresponding weights of each mucosal

Table 2. The performance of EADHI for mucosal features.

Mucosal features	Accuracy % (95% CI)	Sensitivity % (95% CI)	Specificity % (95% CI)	PPV % (95% CI)	NPV % (95% CI)
Atrophy	78.6 (71.7–84.1)	82.3 (72.3–89.3)	75.3 (65.3–83.1)	74.7 (64.6–82.7)	82.7 (72.9–89.5)
Intestinal metaplasia	81.6 (75.0–86.7)	100 (71.8–100)	80.1 (73.1–85.7)	27.9 (16.6–42.8)	100 (96.4–100)
Nodularity	91.7 (86.4–95.1)	71.4 (35.2–92.4)	92.6 (87.3–95.8)	29.4 (13.0–53.4)	98.7 (95.0–99.9)
Mucosal redness	83.3 (76.9–88.3)	79.0 (63.4–89.2)	84.6 (77.4–89.9)	60.0 (46.2–72.4)	93.2 (87.0–96.7)
Mucosal edema	85.7 (79.6–90.3)	87.7 (78.0–93.6)	84.2 (75.5–90.3)	81.0 (70.9–88.3)	89.9 (81.7–94.8)
RAC	82.7 (76.3–87.8)	75.8 (58.8–87.4)	84.4 (77.3–89.7)	54.4 (40.2–67.9)	93.4 (87.4–96.8)
FGP	69.6 (62.3–76.1)	73.3 (47.6–89.5)	69.3 (61.6–76.1)	19.0 (10.8–31.0)	96.4 (90.7–98.9)
Red streak	72.6 (65.4–78.8)	100 (51.1–100)	71.8 (64.4–78.1)	9.8 (3.8–21.4)	100 (96.2–100)
Hematin	58.9 (51.4–66.1)	75.4 (63.2–84.6)	49.5 (40.2–58.9)	46.0 (36.6–55.7)	77.9 (66.6–86.3)
Total	78.3 (76.2–80.3)	81.4 (76.8–85.3)	77.5 (75.0–79.8)	49.5 (45.3–53.8)	93.9 (92.2–95.2)

CI, confidence interval; EADHI, explainable artificial intelligence system for diagnosing *H. pylori* infection; FGP, fundic gland polyp; NPV, negative predictive value; PPV, positive predictive value; RAC, regular arrangement of collecting venules.

feature were 0.109, 0.019, 0.024, 0.163, 0.386, 0.275, 0.012, 0.006, and 0.007, respectively (Figure 7(b)).

Performance of EADHI for H. pylori *infection on the internal test set in patients*

Figure 7(c) shows the AUC of EADHI for *H. pylori* infection was 0.96, and at the optimal threshold of 0.26, the accuracy, sensitivity, and specificity were 91.1% (95% CI: 85.7–94.6%), 92.9% (95% CI: 85.0–97.0%), and 89.3% (95% CI: 80.7–94.5%) in the internal test set, respectively (Table 3).

In the internal test set, 94.0% (158/168) cases were diagnosed by *H. pylori* breath test. EADHI achieved an accuracy of 92.4% (95% CI: 87.1–95.7%), a sensitivity of 93.2% (95% CI: 84.8–97.4%), and a specificity of 91.7% (95% CI: 83.5–96.2%) in these cases (Table 3).

Comparison between EADHI and endoscopists

Table 3 shows the results of *H. pylori* infection evaluation of the internal test data by the 10 endoscopists. The overall accuracy, sensitivity, and specificity for the diagnosis of *H. pylori* infection were 75.6% (SD 8.1%), 78.7% (12.7%), and 72.4% (15.5%). The expert group was found to have significantly higher accuracy (83.6% *versus* 73.2%, $p < 0.01$) than the relatively experienced group. Similarly, a significant difference in the accuracy was observed between the expert group and the beginner group (83.6% *versus* 67.3%, $p < 0.01$). However, there was no statistical difference in the accuracy between the relatively experienced group and the beginner group (73.2% *versus* 67.3%, $p = 0.618$).

In addition, the EADHI was found to have a significantly higher accuracy (by 15.5%; 95% CI: 9.7–21.3%), sensitivity (by 14.2%; 95% CI: 5.1–23.3%), and specificity (by 16.9%; 95% CI: 5.8–28.0%) than the endoscopists. When compared with the expert group, the EADHI had higher accuracy (by 7.5%; 95% CI: 2.5–12.4%), although their sensitivity and specificity were comparable. The comparison results are summarized in Figure 7(c) and Table 3.

Performance of EADHI for H. pylori *infection on the external test set in patients*

To verify the robustness of the EADHI, we examined its performance on an external test set. It achieved an accuracy of 91.9% (95% CI: 85.6–95.7%), a sensitivity of 90.3% (95% CI: 80.1–95.8%), and a specificity of 93.6% (95% CI: 84.1–97.9%), respectively. And the AUC value of 0.96 was achieved at the optimal threshold (Figure 7(c)).

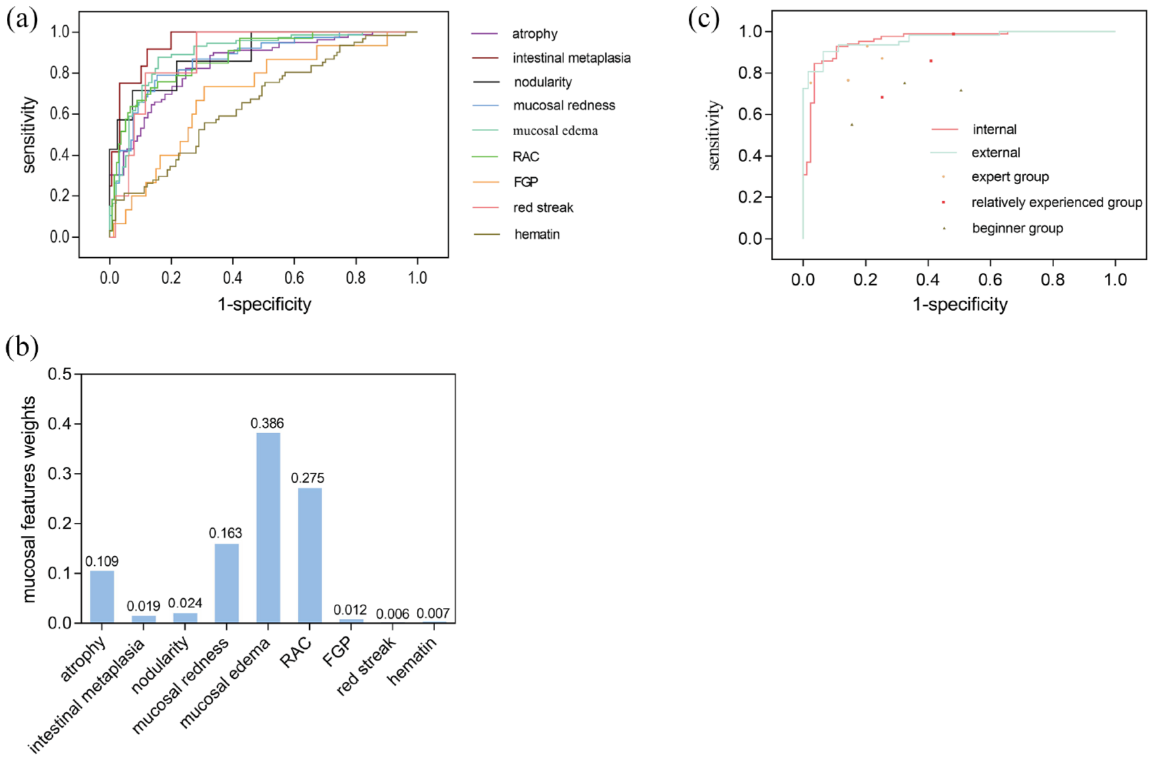

Figure 7. The corresponding weights of each mucosal feature, the ROC curves of EADHI, and the performance of endoscopists. (a) ROC curves of various mucosal features. (b) The corresponding weights of each feature. The ROC curve for *H. pylori* infection in internal testing dataset. (c) The ROC curve for *H. pylori* infection in internal and external testing dataset, and the performance of endoscopists.
EADHI, explainable artificial intelligence system for diagnosing *H. pylori* infection; ROC, receiver operating characteristic.

Table 3. Diagnostic accuracy in internal test data: EADHI *versus* endoscopists.

Index (%)	EADHI		Endoscopists (SD)			
	Cases diagnosed by *H. pylori* breath test	All cases	Expert ($n = 4$)	Relatively experienced ($n = 3$)	Beginner ($n = 3$)	Total ($n = 10$)
Accuracy	92.4	91.1	83.6 (3.1)**	73.2 (2.1)	67.3 (5.7)	75.6 (8.1)
Sensitivity	93.2	92.9	82.9 (8.5)	84.3 (15.4)	67.4 (10.7)	78.7 (12.7)
Specificity	91.7	89.3	84.3 (9.9)*	61.9 (11.7)	67.1 (17.5)	72.4 (15.5)
PPV	90.8	89.7	85.4 (8.2)**	69.8 (3.2)	69.2 (9.6)	75.9 (10.6)
NPV	93.9	92.6	83.5 (6.2)	82.6 (14.1)	66.9 (5.1)	78.2 (11.2)

Comparison between other groups and experts, *indicates $p < 0.05$, **indicates $p < 0.01$.
EADHI, explainable artificial intelligence system for diagnosing *H. pylori* infection; NPV, negative predictive value; PPV, positive predictive value; SD, standard deviation.

Discussion

We developed an explainable AI system, EADHI, for auto-identification of *H. pylori* infection by identifying multiple mucosal features. We compared EADHI's diagnostic ability with endoscopists by examining its internal and external test data performance, and EADHI performed better than that endoscopists and showed good robustness.

In recent years, AI has remarkable improved in diagnosing *H. pylori* infection based on CNN.[25,26] However, because CNN models were developed with a few endoscopic images per patient, mucosal features associated with *H. pylori* infection status may have been missed. In addition, there have been concerns about the black-box nature and lack of explainability of AI models, which greatly limits their clinical application.[27] In this study, two models were integrated into the system to make a diagnosis of EADHI more specific and comprehensive. Model 1 was first constructed with ResNet-50 to enhance its ability to identify mucosal features. In addition, we confirmed which parts of the images model 1 focuses on using heatmaps. Model 2 was developed by combining ResNet-50 and LSTM networks using case-based images (about 26 images per case). Furthermore, ResNet-50 in model 2 incorporated all layers of model 1 except the fully connected layer to maximize the ability of model 2 to recognize mucosal features. In addition, the information on mucosal features in the cases was added during the training to strengthen further the ability of the system to identify mucosal features.

Previous studies show that an estimated 40–50% of the global population is infected with *H. pylori*, which is the main risk factor for GC.[2,37,38] In addition, *H. pylori* infection affects the morphology of EGC and makes it difficult to diagnose.[10–12] Therefore, early detection of *H. pylori* infection using endoscopy is crucial. As more endoscopic findings associated with *H. pylori* infection status are identified, WLE may be used to diagnose *H. pylori* gastritis.[16–19] However, this approach requires advanced skills and knowledge, and the diagnostic process is highly subjective.[20,21] In contrast, an AI-aided diagnosis system could provide an objective second opinion and help endoscopists avoid over-reliance on prior experience in the diagnosis process.[39] Herein, the EADHI developed by us could effectively identify mucosal features and comprehensively detect *H. pylori* infection based on identified features. Our results indicate its potential to assist endoscopists in screening *H. pylori* infection in real clinical work.

Nodularity is considered to be the strongest evidence to support current *H. pylori* infection, with extremely high specificity (95.8–98.8%).[40–42] A recent study reported that the diagnostic odds ratios (DOR) of diffuse redness, mucosal swelling, sticky mucus, and enlarged serpentine for judging *H. pylori* positive were 26.8, 13.3, 10.2, and 8.6, respectively.[42] Moreover, most patients with atrophic gastritis have evidence of *H. pylori* infection.[43,44] In addition, Zhao *et al*. reported that when the positive signs were combined, the ROC/AUC of two or more features had the highest value (0.723).[40] Among these negative mucosal features, the DOR of RAC, FGP, and red streak were 32.2, 7.7, and 4.7, respectively.[42] Furthermore, when these negative features were combined, one or more had the highest ROC/AUC (0.701).[40] These mucosal features make the diagnosis of *H. pylori* infection status easier than before using WLE, but endoscopists' skill levels vary significantly.[22,40] In our study, the expert group performed significantly better than the other groups. Therefore, an effective method for identifying various mucosal features is required. Model 1 was developed in EADHI to improve its ability to identify mucosal features, and it achieved good diagnostic performance.

The EADHI for detecting *H. pylori* infection combines ResNet-50 and LSTM networks, with ResNet-50 being used for feature extraction and LSTM to classify *H. pylori* infection based on these features. The feature extraction network (ResNet-50) is a residual deep learning network (with 50 layers) which attempts to address the problem of vanishing gradients that occur during back-propagation of CNN to effectively extract and recognize the local and global features of images.[45] Furthermore, a recent study reported that ResNet-50 using image generation techniques based on an 80% training set resulted in nearly perfect multi-class prediction accuracy (98.99%) in a 20% validation set.[46] The LSTM network is capable of learning from imperative experiences with long-term states by introducing gate functions into the cell structure. In the case of LSTM, nodes are connected from a directed graph along a time series that is treated as an input with a specific order.[47] Hence, ResNet-50 and LSTM layout feature combination effectively extracted and integrated mucosal features from about 26 endoscopy images per patient, significantly improving the classification.

There are some limitations to this study. First, we invited two doctoral students and two experts to preprocess the data to obtain high-quality labeled data at a low cost. However, it would be better if more endoscopists participated in data preprocessing. Second, the EADHI was developed using data collected retrospectively from a single center and was not prospectively validated. However, we examined EADHI performance in the external testing dataset and found it accurate. Third, we excluded patients with a history of *H. pylori* eradication from the study. Further research should be conducted to identify previous infections. Finally, *H. pylori* infection status was confirmed in most patients using one test. However, *H. pylori* breath test, which has high sensitivity (>95%) and specificity (95%) for the diagnosis of *H. pylori* infection, was used to diagnose >90% of the positive cases and all the negative cases in this study.[48] Furthermore, we excluded patients who received *H. pylori* eradication or had antibiotics within a month or proton pump inhibitor within 2 weeks of *H. pylori* breath test. Thus, the possibility of a false positive or negative *H. pylori* diagnosis was considered insignificant.

In conclusion, the explainable AI system, EADHI, outperformed the endoscopists in diagnostic accuracy. Furthermore, its diagnostic logic is similar to that of endoscopists, which may increase endoscopists' trust and acceptability. It has a high potential for assisting endoscopists in screening for *H. pylori* infection in clinical settings. Further research should be conducted to validate and globally apply the AI-based diagnostic system.

Declarations

Ethics approval and consent to participate

The study was approved by the Ethics Committee of Renmin Hospital of Wuhan University (No: WDRY2022-K056) and Wenzhou Central Hospital (No: L2022-04-055), and conducted under the Declaration of Helsinki. The Institutional Review Board waived informed consent for retrospective image data.

Consent for publication

Not applicable.

Author contribution(s)

Mengjiao Zhang: Data curation; Methodology; Writing – original draft.

Jie Pan: Data curation; Investigation; Resources; Writing – original draft.

Jiejun Lin: Data curation; Investigation; Writing – review & editing.

Ming Xu: Data curation; Formal analysis; Writing – review & editing.

Lihui Zhang: Data curation; Investigation; Writing – review & editing.

Renduo Shang: Data curation; Investigation; Writing – review & editing.

Liwen Yao: Data curation; Investigation; Writing – review & editing.

Yanxia Li: Data curation; Formal analysis; Writing – review & editing.

Wei Zhou: Data curation; Formal analysis; Writing – review & editing.

Yunchao Deng: Data curation; Formal analysis; Writing – review & editing.

Zehua Dong: Data curation; Investigation; Writing – review & editing.

Yijie Zhu: Data curation; Investigation; Writing – review & editing.

Xiao Tao: Data curation; Investigation; Writing – review & editing.

Lianlian Wu: Data curation; Formal analysis; Funding acquisition; Methodology; Writing – review & editing.

Honggang Yu: Conceptualization; Funding acquisition; Project administration; Resources; Writing – review & editing.

Acknowledgements

We thank the fund support from Innovation Team Project of Health Commission of Hubei Province (Grant no. WJ202C003); The Fundamental Research Funds for the Central Universities (Grant no. 2042021kf0084). We thank the endoscopists and machine-learning engineers for their hard work.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Innovation Team Project of Health Commission of Hubei Province (Grant no.WJ202C003); The Fundamental Research Funds for the Central Universities (Grant no. 2042021kf0084).

Competing interests

The authors declare that there is no conflict of interest.

Availability of data and materials

Individual de-identified participant data that underlie the results reported in this article and study protocol will be shared for investigators after article publication. To gain access, data requesters will need to contact the corresponding author.

ORCID iD

Honggang Yu https://orcid.org/0000-0001-5986-1284

Supplemental material

Supplemental material for this article is available online.

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. Plummer M, Franceschi S, Vignat J*, et al.* Global burden of gastric cancer attributable to *Helicobacter pylori*. *Int J Cancer* 2015; 136: 487–490.
3. Take S, Mizuno M, Ishiki K*, et al.* Seventeen-year effects of eradicating *Helicobacter pylori* on the prevention of gastric cancer in patients with peptic ulcer; a prospective cohort study. *J Gastroenterol* 2015; 50: 638–644.
4. Shichijo S, Hirata Y, Sakitani K*, et al.* Distribution of intestinal metaplasia as a predictor of gastric cancer development. *J Gastroenterol Hepatol* 2015; 30: 1260–1264.
5. Shichijo S, Hirata Y, Niikura R*, et al.* Histologic intestinal metaplasia and endoscopic atrophy are predictors of gastric cancer development after *Helicobacter pylori* eradication. *Gastrointest Endosc* 2016; 84: 618–624.
6. Wang F, Meng W, Wang B*, et al. Helicobacter pylori*-induced gastric inflammation and gastric cancer. *Cancer Lett* 2014; 345: 196–202.
7. Tahara T, Shibata T, Horiguchi N*, et al.* A possible link between gastric mucosal atrophy and gastric cancer after *Helicobacter pylori* eradication. *PLoS One* 2016; 11: e0163700.
8. Valori R, Cortas G, de Lange T*, et al.* Performance measures for endoscopy services: a European Society of Gastrointestinal Endoscopy (ESGE) quality improvement initiative. *United European Gastroenterol J* 2019; 7: 21–44.
9. Yao K, Uedo N, Kamada T*, et al.* Guidelines for endoscopic diagnosis of early gastric cancer. *Dig Endosc* 2020; 32: 663–698.
10. Mitsuhashi J, Mitomi H, Tanabe S*, et al.* Differences in clinicopathological findings, cell kinetics and p53 expression between early gastric cancers with and without *Helicobacter pylori* infection. *Hepatogastroenterology* 2004; 51: 1636–1640.
11. Sato C, Hirasawa K, Tateishi Y*, et al.* Clinicopathological features of early gastric cancers arising in *Helicobacter pylori* uninfected patients. *World J Gastroenterol* 2020; 26: 2618–2631.
12. Pasechnikov V, Chukov S, Fedorov E*, et al.* Gastric cancer: prevention, screening and early diagnosis. *World J Gastroenterol* 2014; 20: 13842–13862.
13. Suzuki H and Moayyedi P. *Helicobacter pylori* infection in functional dyspepsia. *Nat Rev Gastroenterol Hepatol* 2013; 10: 168–174.
14. Lee SP, Lee J, Kae SH*, et al.* The role of linked color imaging in endoscopic diagnosis of *Helicobacter pylori* associated gastritis. *Scand J Gastroenterol* 2020; 55: 1114–1120.
15. Wang L, Lin XC, Li HL*, et al.* Clinical significance and influencing factors of linked color imaging technique in real-time diagnosis of active *Helicobacter pylori* infection. *Chin Med J* 2019; 132: 2395–2401.
16. Kamada T, Haruma K, Inoue K*, et al.* [*Helicobacter pylori* infection and endoscopic gastritis–Kyoto classification of gastritis]. *Nihon Shokakibyo Gakkai Zasshi* 2015; 112: 982–993.
17. Kato M. *Endoscopic findings of H. pylori infection* In: Suzuki H, Warren R and Marshall B (eds) *Helicobacter pylori*. Tokyo: Springer, 2016, pp.157–167.
18. Glover B, Teare J, Ashrafian H*, et al.* The endoscopic predictors of *Helicobacter pylori* status: a meta-analysis of diagnostic performance. *Ther Adv Gastrointest Endosc* 2020; 13: 2631774520950840.
19. Glover B, Teare J and Patel N. Assessment of *Helicobacter pylori* status by examination of gastric mucosal patterns: diagnostic accuracy of white-light endoscopy and narrow-band imaging. *BMJ Open Gastroenterol* 2021; 8: e000608.
20. Sugano K, Tack J, Kuipers EJ*, et al.* Kyoto global consensus report on *Helicobacter pylori* gastritis. *Gut* 2015; 64: 1353–1367.
21. Watanabe K, Nagata N, Shimbo T*, et al.* Accuracy of endoscopic diagnosis of *Helicobacter pylori* infection according to level of endoscopic

experience and the effect of training. *BMC Gastroenterol* 2013; 13: 128.

22. Anwar SM, Majid M, Qayyum A, *et al*. Medical image analysis using convolutional neural networks: a review. *J Med Syst* 2018; 42: 226.
23. Hashimoto R, Requa J, Dao T, *et al*. Artificial intelligence using convolutional neural networks for real-time detection of early esophageal neoplasia in Barrett's esophagus (with video). *Gastrointest Endosc* 2020; 91: 1264–1271.e1.
24. Wu L, Xu M, Jiang X, *et al*. Real-time artificial intelligence for detecting focal lesions and diagnosing neoplasms of the stomach by white-light endoscopy (with videos). *Gastrointest Endosc* 2021; 95: 269–280.e6.
25. 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: 106–111.
26. 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.
27. Holzinger A, Langs G, Denk H, *et al*. Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip Rev Data Min Knowl Discov* 2019; 9: e1312.
28. Zhang Y, Weng Y and Lund J. Applications of explainable artificial intelligence in diagnosis and surgery. *Diagnostics (Basel)* 2022; 12: 237.
29. Kundu S. AI in medicine must be explainable. *Nat Med* 2021; 27: 1328.
30. Stoltzfus JC. Logistic regression: a brief primer. *Acad Emerg Med* 2011; 18: 1099–1104.
31. Shorten C and Khoshgoftaar TM. A survey on image data augmentation for deep learning. *J Big Data* 2019; 6: 60.
32. Selvaraju RR, Cogswell M, Das A, *et al*. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int J Comput Vis* 2020; 128: 336–359.
33. Mohammad N, Muad AM, Ahmad R, *et al*. Accuracy of advanced deep learning with tensorflow and keras for classifying teeth developmental stages in digital panoramic imaging. *BMC Med Imaging* 2022; 22: 66.
34. Li T, Yang K, Stein JD, *et al*. Gradient boosting decision tree algorithm for the prediction of postoperative intraocular lens position in cataract surgery. *Transl Vis Sci Technol* 2020; 9: 38.
35. Ren S, Cai P, Liu Y, *et al*. Prevalence of *Helicobacter pylori* infection in China: a systematic review and meta-analysis. *J Gastroenterol Hepatol* 2022; 37: 464–470.
36. von Elm E, Altman DG, Egger M, *et al*. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Int J Surg* 2014; 12: 1495–1499.
37. Lee YC, Chiang TH, Chou CK, *et al*. Association between *Helicobacter pylori* eradication and gastric cancer incidence: a systematic review and meta-analysis. *Gastroenterology* 2016; 150: 1113–1124.e5.
38. Hooi JKY, Lai WY, Ng WK, *et al*. Global prevalence of *Helicobacter pylori* infection: systematic review and meta-analysis. *Gastroenterology* 2017; 153: 420–429.
39. Tang Y, Anandasabapathy S and Richards-Kortum R. Advances in optical gastrointestinal endoscopy: a technical review. *Mol Oncol* 2021; 15: 2580–2599.
40. Zhao J, Xu S, Gao Y, *et al*. Accuracy of endoscopic diagnosis of *Helicobacter pylori* based on the Kyoto classification of gastritis: a multicenter study. *Front Oncol* 2020; 10: 599218.
41. Kato T, Yagi N, Kamada T, *et al*. Diagnosis of *Helicobacter pylori* infection in gastric mucosa by endoscopic features: a multicenter prospective study. *Dig Endosc* 2013; 25: 508–518.
42. Yoshii S, Mabe K, Watano K, *et al*. Validity of endoscopic features for the diagnosis of *Helicobacter pylori* infection status based on the Kyoto classification of gastritis. *Dig Endosc* 2020; 32: 74–83.
43. Oksanen A, Sipponen P, Karttunen R, *et al*. Atrophic gastritis and *Helicobacter pylori* infection in outpatients referred for gastroscopy. *Gut* 2000; 46: 460–463.
44. Annibale B, Negrini R, Caruana P, *et al*. Two-thirds of atrophic body gastritis patients have evidence of *Helicobacter pylori* infection. *Helicobacter* 2001; 6: 225–233.
45. He K, Zhang X, Ren S, *et al*. *Deep residual learning for image recognition* In: *2016 IEEE conference on computer vision and pattern recognition (CVPR)* 2016; pp.770–778. New York: IEEE.
46. Fulton LV, Dolezel D, Harrop J, *et al*. Classification of Alzheimer's disease with and without imagery using gradient boosted machines and ResNet-50. *Brain Sci* 2019; 9: 212.
47. Yu Y, Si X, Hu C, *et al*. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput* 2019; 31: 1235–1270.
48. Huh CW and Kim BW. [Diagnosis of *Helicobacter pylori* Infection]. *Korean J Gastroenterol* 2018; 72: 229–236.

Visit SAGE journals online
journals.sagepub.com/
home/tag

SAGE journals