BMC Medicine

RESEARCH

Open Access

Artificial intelligence–assisted endoscopic diagnosis system for diagnosing *Helicobacter pylori* infection: a multicenter study

Yue Hu[1,2†], Jianwei Xu[3†], Liang Huang[1,2†], Zhilin Zheng[3], Jin Zhao[1,2], Tanzhou Chen[4], Jiang Liu[5], Fangfang Zhang[6], Xiaoyun Ding[7], Jie Pan[8], Xiaoteng Wang[9], Shan Liu[10], Ling Zhang[3*] and Bin Lv[1,2*]

Abstract

Background Deep learning algorithm-based artificial intelligence (AI) has significantly advanced the domain of endoscopic diagnosis; however, its utilization for detecting *Helicobacter pylori* (*H. pylori*) infections remains constrained. We aimed to develop and validate the AI diagnostic system (HOPE AI) for diagnosing *H. pylori* infection by analyzing extensive imaging data obtained from clinical endoscopies.

Methods This multicenter diagnostic study was carried out across seven hospitals in China. Eligible patients were individuals aged 18 years or older who underwent upper gastrointestinal gastroendoscopy. The endoscopic images were randomly allocated (7:3) to the training and internal validation datasets for the development of HOPE AI, utilizing a multi-instance learning (MIL) framework and long short-term memory (LSTM) architectures, and the prospective external validation dataset for assessing its diagnostic efficacy. The performance of HOPE AI was also benchmarked against endoscopists. The diagnostic accuracy, sensitivity, specificity, and area under the curve of HOPE AI were assessed to detect *H. pylori* infection.

Results A total of 308,887 endoscopic images and 197 videos from 6207 patients were utilized to develop and evaluate HOPE AI. Our AI system demonstrated outstanding performance, achieving an AUC of 0.932 (95% confidence interval (CI) 0.906–0.956) in the internal validation set, 0.903 (0.883–0.922) in the external temporal validation set, 0.923 (0.875–0.961) in the external temporal validation video set, and ranging from 0.855 (0.813–0.894) to 0.971 (0.955–0.985) across seven external geographical validation sets. The diagnostic sensitivity of HOPE AI (85.7%) significantly surpassed that of senior endoscopists (68.0%).

Conclusions HOPE AI exhibited robust diagnostic efficacy and interpretability in *H. pylori* detection, thereby enhancing the efficiency of diagnosis in routine screening contexts.

Trial registration Chinese Clinical Trial Registry: ChiCTR 2400091317, 2,400,091,720.

Keywords *Helicobacter pylori*, Endoscopy, Artificial intelligence, Multi-instance learning

†Yue Hu, Jianwei Xu and Liang Huang contributed equally to this work.

*Correspondence:
Ling Zhang
ling.z@alibaba-inc.com
Bin Lv
lvbin@medmail.com.cn
Full list of author information is available at the end of the article

Background

Helicobacter pylori (*H. pylori*), classified as a Group I carcinogenic agent by the World Health Organization (WHO) and the International Agency for Research on Cancer (IARC) [1, 2], initiates inflammatory cascades in the gastric mucosa, manifesting as both acute and chronic gastritis. This persistent inflammation can precipitate a sequential pathological progression encompassing gastric mucosal atrophy, intestinal metaplasia, dysplasia, and, ultimately, the development of gastric cancer. Therefore, in current clinical practice, prompt diagnosis and eradication of *H. pylori* infection represent the most efficacious strategy in mitigating gastric carcinoma risk [3].

Gastroendoscopic evaluation represents one of the critical diagnostic modalities in identifying *H. pylori* infection. Endoscopic findings can support a diagnosis of *H. pylori* infection based on the representative features according to the Kyoto classification score [4] and the modified system [5]. The character endoscopic features indicative of *H. pylori* infection encompass gastric mucosal nodularity, diffuse redness, atrophy, or mucosal swelling. Conversely, the presence of regular arrangement of collecting venules (RAC) and fundic gland polyps strongly suggests *H. pylori*-negative status, making it possible to diagnose accurately endoscopically. Nevertheless, the field currently lacks standardized objective visualization parameters for *H. pylori* identification via endoscopy, with significant heterogeneity in diagnostic accuracy across healthcare institutions and clinicians of varying expertise levels [6], presenting substantial impediments to advancement in this domain.

Leveraging the inherent capabilities of computer vision and pattern recognition, artificial intelligence (AI) systems powered by deep learning algorithms have demonstrated remarkable advancement in endoscopic diagnostic applications [7]. The technology has shown particular efficacy in discriminating between neoplastic and non-neoplastic lesions, representing the most predominant clinical implementation of AI-assisted endoscopy [8, 9]. Recent researchers have expanded the scope of AI applications to *H. pylori* infection detection, broadening its utility in gastroenterological diagnostics [10–13]. While initial studies investigating AI-based *H. pylori* detection systems have yielded promising results, their clinical applications remain constrained by several methodological limitations, including insufficient validation cohorts, retrospective analytical approaches, single-institution study designs, and other constraints.

In this study, we present an innovative approach for developing HOPE AI (simplified from *Helicobacter pylori* identification AI), a visualized, interpretable, and stable AI system designed for *H. pylori* infection diagnosis. The system is based on a multi-instance learning framework (MIL) and incorporates transformer and long short-term memory (LSTM) architectures. We evaluated HOPE AI's diagnostic efficacy through comprehensive internal and external validation using a substantial multicenter endoscopic dataset collected from seven medical institutions. This represents the first large-scale external validation study of its kind, demonstrating the model's generalizability and establishing a strong foundation for clinical implementation.

Methods

Study design

This multi-institutional diagnostic investigation was implemented across seven Chinese medical centers, encompassing three distinct methodological phases: (phase 1) development and internal validation of the HOPE AI algorithm utilizing clinical endoscopic imagery from the First Affiliated Hospital of Zhejiang Chinese Medical University, Hubin Campus (FAHZCMU, H); (phase 2) external temporal validation employing prospectively acquired endoscopic images and video sequences from FAHZCMU, H; and (phase 3) external geographic validation through prospective data collection from multiple institutions, including FAHZCMU Qiantang Campus and six additional regional hospitals (the First Affiliated Hospital of Wenzhou Medical University, the First Hospital of Jiaxing, Wenzhou Central Hospital, the First Affiliated Hospital of Ningbo University, Huzhou Central Hospital, and Yuyao People's Hospital of Zhejiang). The research protocol adhered to Helsinki Declaration guidelines and received institutional review board approval from FAHZCMU (No. 2024-KLS-489–01), agreed by other centers. The study was registered in the Chinese Clinical Trial Registry (ChiCTR 2400091317, 2,400,091,720).

Study population

In this multicenter ambispective cohort study, the retrospective derivation cohort (phase 1) encompassed 1271 patients during the period from March 1 to June 30, 2024. Subjects were stratified through randomization (7:3 ratio) into training and internal validation datasets for HOPE AI development. The external validation cohort (phases 2 and 3) prospectively recruited 4936 patients across seven medical centers between October 25 and November 30, 2024 (Fig. 1) [14].

Study eligibility criteria were defined as follows: adult patients (≥18 years), regardless of gender, completion of standardized gastroscopic examination (minimum 8 gastric images, including cardia, upper body, middle body, lesser curvature, angle, lower body, antrum, and pylorus [15]), and confirmed *H. pylori* status through

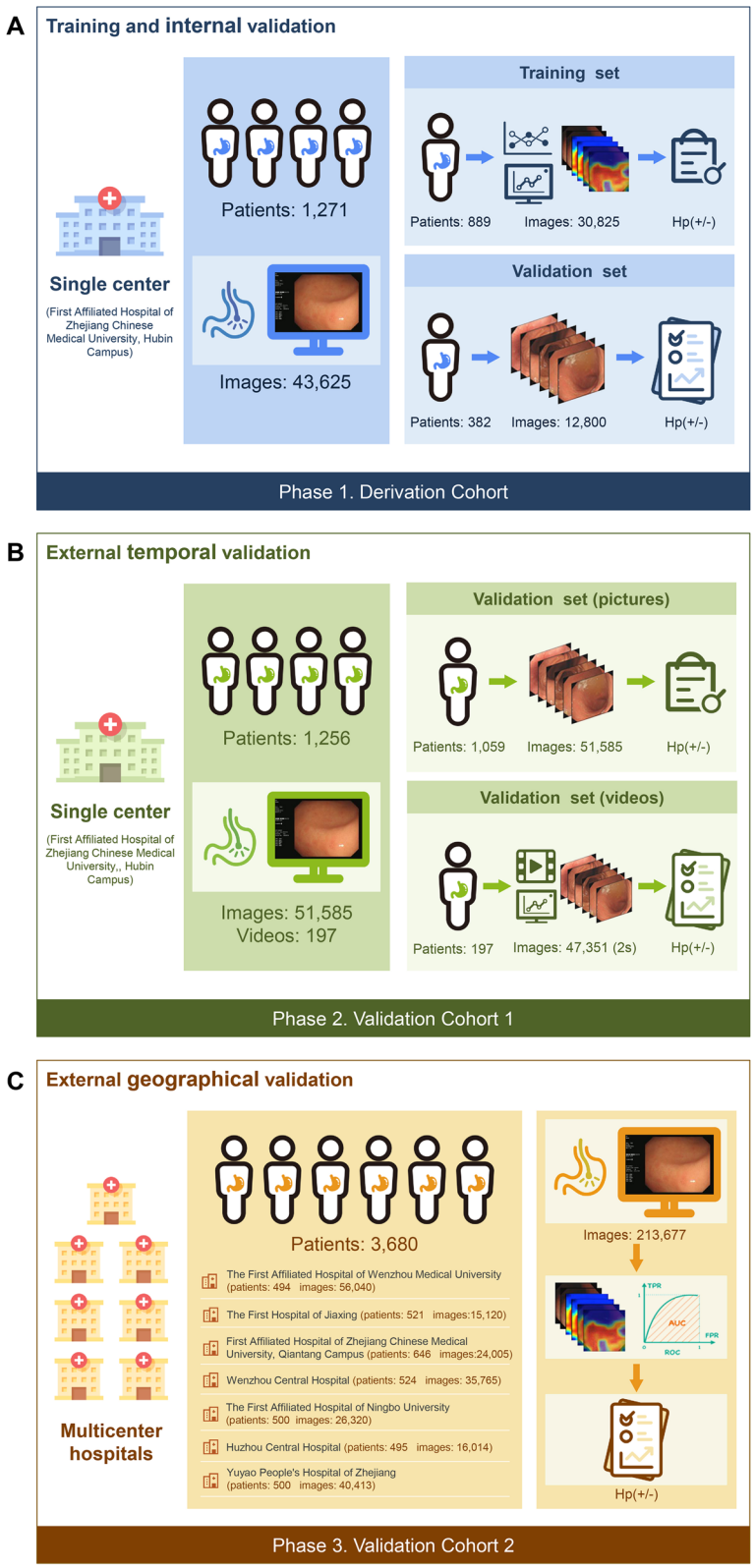

Fig. 1 Overview of the study design. **A** Training and internal validation (phase 1). **B** External temporal validation (phase 2). **C** External geographical validation (phase 3)

histopathological assessment (negative, encompassing both absence of prior *H. pylori* infection and successful previous eradication therapy; positive, indicating current active *H. pylori* infection, regardless of prior anti-*H. pylori* history). For video analyses, an additional criterion mandated complete gastroscopic examination documentation. Exclusion criteria comprised presence of retained gastric contents, neoplastic lesions, luminal obstruction, and documented history of esophagogastric malignancy or gastric surgical intervention.

Development of HOPE AI

In our comprehensive dataset compilation, we incorporated all endoscopic images acquired during routine clinical procedures, ensuring an unbiased representation of real clinical scenarios. The dataset incorporates diverse imaging modalities, including conventional white-light endoscopy (WLE) and narrow-band imaging (NBI), encompassing varying image qualities such as focused, unfocused, overexposed, motion-blurred, and hemorrhagic specimens. The dataset was dichotomously classified based on *Helicobacter pylori* infection status: non-infected (label 0) and infected (label 1). We preprocessed the images by excluding the metadata sidebar, preserving only the diagnostically relevant endoscopic field. The endoscopic images, originally ranging from 400×480 to 1920×1080 pixels, were standardized to 352×352 pixels to facilitate deep learning network compatibility.

In this study, we introduce a novel MIL architecture integrating vision transformer (ViT) [16, 17] and LSTM networks to comprehensively analyze case-wide image aggregation for *H. pylori* infection detection. The MIL framework, trained exclusively on patient-level annotations, autonomously identifies intrinsic image correlations and patterns. This methodology effectively isolates sentinel images for case classification, maintaining robustness against incomplete or noisy datasets. The inference pipeline generates patient-level infection status predictions and identifies infection-associated high-risk images from extensive case repositories, enhancing clinical decision-making precision. The architecture comprises three main components: a transformer-based feature extraction module that computes *H. pylori* risk scores for individual images, a selection mechanism that identifies the top-k high-risk images based on MIL principles, and an LSTM network that synthesizes features from these top-k images to generate case-level infection probability (Fig. 2). All quantitative evaluations in the analysis relied on case-level infection probabilities as the primary metric, and we also analyzed the contribution of image-level predictions to clinical diagnosis. Detailed methodological protocols and implementation details are summarized in the Additional file 1: Supplementary Materials [10, 16–27].

Validation of HOPE AI

We initially evaluated HOPE AI's performance in detecting *H. pylori* infection through an internal validation cohort, followed by a prospective external temporal validation dataset comprising endoscopic imagery and video sequences from FAHZCMU, H. Subsequently, we conducted a comprehensive assessment of HOPE AI's reliability using geographically diverse external validation datasets from seven institutions, each characterized by substantial endoscopic examination volumes and highly credentialed endoscopists.

To further validate and extend the applicability of our AI model, we selected endoscopic video sequences from the prospective external temporal validation cohort due to its enhanced objectivity and mitigation of the selection bias inherent in physician-dependent image acquisition. Three junior endoscopists and three senior endoscopists, blinded to patient demographics and histopathological outcomes, independently analyzed identical test videos, with their assessments compared against HOPE AI's diagnostic output. These junior practitioners possessed less than 2 years of endoscopic experience; the seniors had over 10 years of experience in endoscopic procedures and had conducted more than 5000 examinations. To maintain objectivity, these evaluators were excluded from image selection and annotation processes, and all visual data was mixed and deidentified prior to their assessment.

Statistical analysis

The diagnostic performance metrics of HOPE AI in detecting *H. pylori* infection were comprehensively assessed through statistical analysis of accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV), with 95% confidence intervals computed via the Clopper-Pearson methodology. Statistical comparisons of diagnostic parameters (sensitivity, specificity, and accuracy) were performed via permutation testing, generating two-sided *P* values based on 10,000 permutations. The discriminative capability of the deep learning algorithm was evaluated through receiver operating characteristic (ROC) curve analysis, wherein sensitivity was plotted against 1-specificity across varying probability thresholds. The AUC metric served as a quantitative measure of diagnostic efficacy, with higher values indicating superior performance. All analyses employed two-tailed statistical tests with $\alpha = 0.05$, implemented using Python 3.9.0.

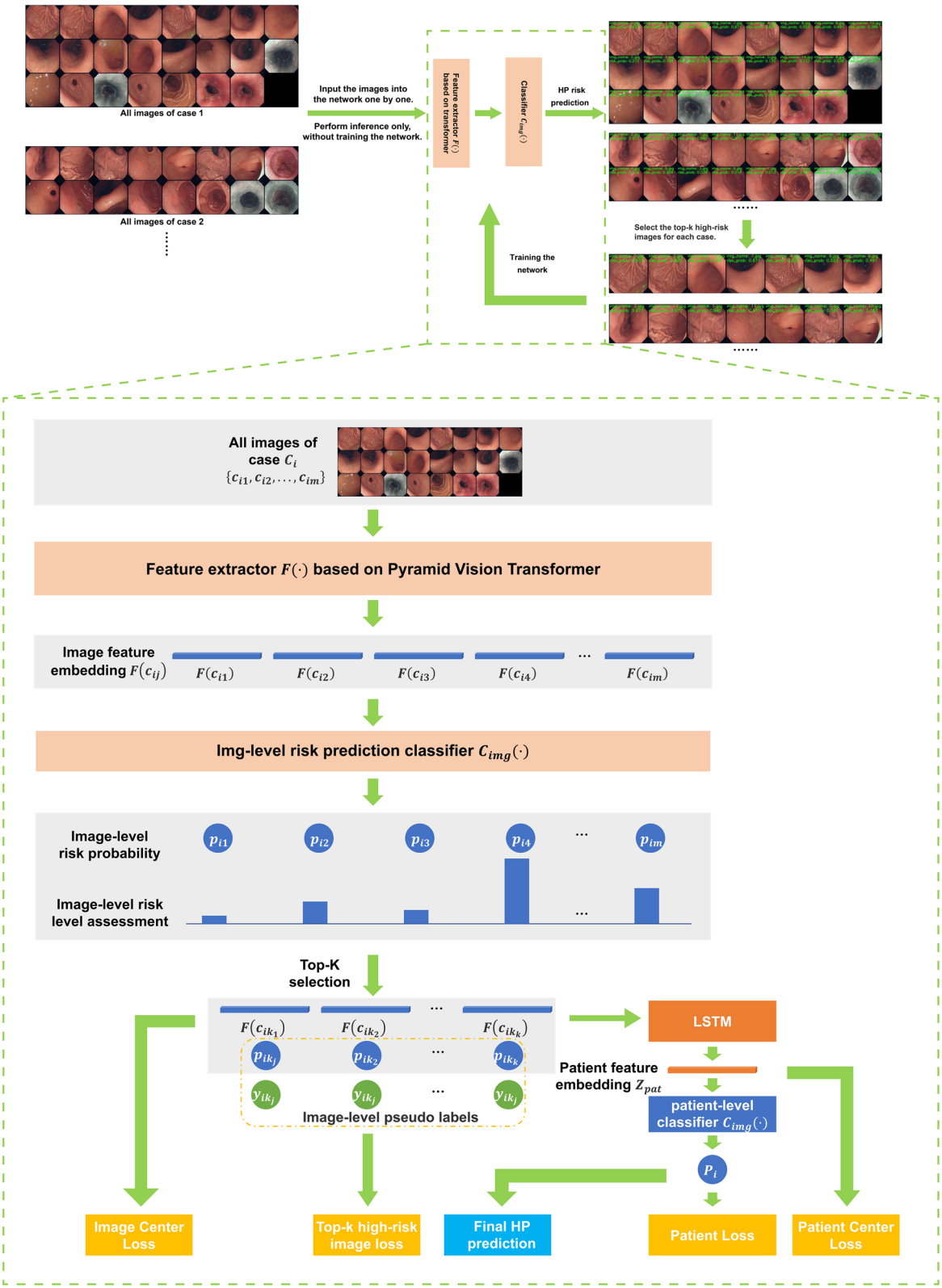

Fig. 2 The development procedure of HOPE AI with TransMIL + LSTM model. The model uses transformers to extract features and classify *H. pylori* infection probability from endoscopic images. It selects top-k high-risk images for LSTM-based feature integration, optimizing through gradient descent with both image and patient-level loss functions. Trans, transformers; MIL, multi-instance learning; LSTM, long short-term memory

Results

Dataset characteristics

During the period from March 1 to June 30, 2024, we collected 43,625 upper gastrointestinal endoscopic images from 1271 subjects at FAHZCMU, H, comprising 473 *H. pylori*-positive and 798 *H. pylori*-negative cases. We allocated 30,825 images from 889 patients (329 positive, 560 negative) to the training dataset, with the remainder assigned to the internal dataset (phase 1). Our prospective external validation dataset included 51,585 images from 1059 patients (366 positive, 693 negative), while the video dataset contained 47,351 images/2 s from 197 patients (49 positive, 148 negative) at FAHZCMU, H (phase 2).

For external geographical validation (phase 3), we prospectively acquired 213,677 endoscopic images from 3680 patients across seven centers between October 25 and November 30, 2024. In total, we utilized 356,238 endoscopic images from 6207 subjects for HOPE AI development and validation (Additional file 2: Table S1). The Phase 1 cohort demonstrated a mean age of 51.2 ± 13.6 years with 44.2% male representation. Phase 2 subjects averaged 54.3 ± 13.8 years with 46.5% males, while phase 3 participants averaged 50.5 ± 14.2 years with 47.8% male distribution. Detailed demographic characteristics are presented in Table 1.

Performance of HOPE AI

The AI system demonstrated robust performance in detecting *H. pylori* infection across multiple validation cohorts (Table 2 and Fig. 3). In the internal testing set, HOPE AI exhibited notable diagnostic metrics, achieving 84.7% (95% CI 78.5–90.3%) sensitivity, 85.7% (81.4–91.2%) specificity, 85.3% (81.68–88.74%) accuracy, and an AUC of 0.932 (0.906–0.956). The external temporal validation using endoscopic imagery maintained comparable efficacy, with 88.0% (84.4–91.2%) sensitivity, 81.53% (78.6–84.3%) specificity, 83.8% (84.5–85.9%) accuracy, and an AUC of 0.903 (0.883–0.922). Video cohort validation yielded an AUC of 0.923 (0.875–0.961), with corresponding sensitivity, specificity, and accuracy values of 85.7%, 85.1%, and 85.3%, respectively.

External geographical validation across multiple medical centers demonstrated consistent performance, with overall metrics of 83.4% (81.3–85.4%) sensitivity, 83.3% (81.8–84.7%) specificity, 83.3% (82.0–84.5%) accuracy, and an AUC of 0.882 (0.870–0.894). Individual institutional performance varied, with AUC values ranging from 0.855 to 0.971 and diagnostic accuracy between 78.9 and 91.2% across seven participating hospitals (Table 2 and Fig. 3).

Notably, in a comparative analysis involving 197 video cases, HOPE AI significantly outperformed endoscopists, both non-expert and expert, demonstrating superior sensitivity (85.7 vs. 54.8%), comparable specificity (85.1 vs. 84.0%), and enhanced accuracy (85.3 vs. 76.7%) in identifying *H. pylori* infection (Additional file 2: Table S2). In further subgroup analyses, HOPE AI did better than the mean performance of the senior endoscopists by 17.7% (95% CI 7.48–28.8%, $p = 0.0068$) in sensitivity while

Table 1 Baseline characteristics

Characteristics	Age, years	Sex		Helicobacter pylori status	
		Male	Female	Positive	Negative
Training (*n* = 889)	50.7 (13.6)	401 (45.1%)	488 (54.9%)	329 (37.0%)	560 (63.0%)
Internal (*n* = 382)	52.4 (13.5)	161 (42.1%)	221 (57.1%)	144 (37.7%)	238 (62.3%)
External temporal validation (*n* = 1256)	54.3 (13.8)	584 (46.5%)	672 (53.5%)	415 (33.0%)	841 (67.0%)
Pictures (*n* = 1059)	53.9 (13.7)	500 (47.2%)	559 (52.8%)	366 (34.6%)	693 (65.4%)
Videos (*n* = 197)	56.2 (14.6)	84 (42.6%)	113 (57.4%)	49 (24.9%)	148 (75.1%)
External geographical validation (*n* = 3680)	50.5 (14.2)	1759 (47.8%)	1921 (52.2%)	1241 (33.7%)	2439 (66.3%)
FAHWMU (*n* = 494)	51.1 (13.6)	244 (49.4%)	250 (50.6%)	105 (21.3%)	389 (78.7%)
FHJ (*n* = 521)	50.1 (14.2)	247 (47.4%)	274 (52.6%)	133 (25.5%)	388 (74.5%)
FAHZCMU, Q (*n* = 646)	46.2 (15.4)	334 (51.7%)	312 (48.3%)	255 (39.5%)	391 (60.5%)
WCH (*n* = 524)	55.3 (12.6)	248 (47.3%)	275 (52.7%)	196 (37.4%)	328 (62.6%)
FAHNU (*n* = 500)	47.9 (15.1)	229 (45.8%)	271 (54.2%)	184 (36.8%)	316 (63.2%)
HCH (*n* = 495)	50.6 (13.4)	233 (47.1%)	262 (52.9%)	202 (40.8%)	293 (59.2%)
YPH (*n* = 500)	53.7 (12.6)	223 (44.6%)	277 (55.4%)	166 (33.2%)	334 (66.8%)

Data are mean (SD) or *n* (%)

FAHWMU The First Affiliated Hospital of Wenzhou Medical University, *FHJ* The First Hospital of Jiaxing, *FAHZCMU, Q* The First Affiliated Hospital of Zhejiang Chinese Medical University, Qiantang Campus, *WCH* Wenzhou Central Hospital, *FAHNU* The First Affiliated Hospital of Ningbo University, *HCH* Huzhou Central Hospital, *YPH* Yuyao People's Hospital of Zhejiang

Table 2 Performance of HOPE AI in different validation sets

	Specificity (95% CI)	Sensitivity (95% CI)	Accuracy (95% CI)	Positive predictive value (95% CI)	Negative predictive value (95% CI)
Internal (*n* = 382)	85.71 (81.09–89.92)	84.72 (78.47–90.28)	85.34 (81.68–88.74)	78. 21 (72.84–83.67)	90. 27 (86.84–93.61)
External temporal validation (*n* = 1256)					
Pictures (*n* = 1059)	81.53 (78.64–4.27)	87.98 (84.43–91.26)	83.76 (81.49–85.93)	71. 56 (68.37–74.88)	92. 78 (90.86–94.62)
Videos (*n* = 197)	85.14 (79.05–90.54)	85.71 (75.51–93.88)	85.28 (80.20–90.36)	65.63 (56.96–75.47)	94.74 (91.24–97.86)
External geographical validation (*n* = 3680)	83.27 (81.75–84.71)	83.40 (81.29–85.40)	83.32 (82.03–84.48)	71.73 (69.80–73.55)	90.79 (89.72–91.82)
FAHWMU (*n* = 494)	80.46 (76.35–84.32)	83.81 (76.19–90.48)	81.17 (77.73–84.62)	53.66 (48.57–59.35)	94.85 (92.65–96.93)
FHJ (*n* = 521)	79.90 (75.77–83.76)	75.94 (68.42–82.71)	78.89 (75.43–82.34)	56.42 (51.23–61.99)	90.64 (88.02–93.20)
FAHZCMU, Q (*n* = 646)	85.68 (82.10–89.00)	87.45 (83.14–91.37)	86.38 (83.75–89.98)	79.93 (76.03–83.97)	91.28 (88.68–93.84)
WCH (*n* = 524)	81.71 (77.13–85.37)	86.73 (81.54–91.28)	83.59 (80.11–86.42)	73.91 (69.17–78.10)	91.16 (88.14–93.97)
FAHNU (*n* = 500)	81.01 (76.58–85.13)	78.26 (72.28–84.24)	80.00 (76.40–83.60)	70.59 (65.69–75.77)	86.49 (83.23–89.80)
HCH (*n* = 495)	83.96 (79.52–88.05)	77.23 (71.29–82.67)	81.21 (77.78–84.44)	76.85 (72.06–81.77)	84.25 (80.84–87.68)
YPH (*n* = 500)	90.72 (87.43–93.71)	92.17 (87.95–95.78)	91.20 (88.60–93.60)	83.15 (78.57–87.86)	95.89 (93.75–97.78)

FAHWMU The First Affiliated Hospital of Wenzhou Medical University, *FHJ* The First Hospital of Jiaxing, *FAHZCMU, Q* The First Affiliated Hospital of Zhejiang Chinese Medical University, Qiantang Campus, *WCH* Wenzhou Central Hospital, *FAHNU* The First Affiliated Hospital of Ningbo University, *HCH* Huzhou Central Hospital, *YPH* Yuyao People's Hospital of Zhejiang

maintaining comparable specificity (difference of 1.1%, 95% CI − 2.5–4.7%, $p = 0.7984$) (Supplementary Table 2). Compared with junior endoscopists, the performance differential was even more pronounced, with a 44.2% (95% CI 33.9–54.5%, $p < 0.0001$) higher sensitivity while maintaining similar specificity (difference of 1.1%, 95% CI − 2.1–4.3%, $p = 0.7384$) (Additional file 2: Table S3).

Explainability of HOPE AI

Through visual analysis at the image level, we demonstrated comparative endoscopic images of *H. pylori*-positive and *H. pylori*-negative cases, accompanied by risk probability assessments generated by the MIL algorithm and expert gastroenterologists' clinical feature annotations for individual images (Fig. 4A). Additionally, we present characteristic attention maps highlighting regions of *H. pylori* colonization at the pixel-level resolution, demonstrating the focal points of our artificial intelligence algorithms (Fig. 4B–F). Notably, the regions of interest identified through our computational heat mapping demonstrated significant concordance with suspicious areas independently identified by experienced endoscopists. These computer-generated focal regions represent clinically significant predictive features, thereby validating the interpretability of our HOPE AI diagnostic system.

Discussion

Our research presents HOPE AI, an innovative artificial intelligence-driven diagnostic system for *H. pylori* detection, developed using multiple-instance learning methodology. This comprehensive system was developed and validated utilizing an extensive dataset comprising 308,887 endoscopic images and 197 video sequences, collected from 6207 subjects across seven medical institutions with varying expertise levels in *H. pylori* infection diagnosis. The system demonstrated exceptional diagnostic performance metrics, including robust accuracy, sensitivity, and specificity, both in retrospective image analysis and prospective clinical evaluation. This investigation represents the most extensive validated cohort study globally in the domain of AI-assisted *H. pylori* detection utilizing upper gastrointestinal endoscopic imaging.

H. pylori infection, a well-documented etiological factor in gastric carcinogenesis, demonstrates global prevalence exceeding 50% [28]. While *H. pylori* eradication represents a crucial therapeutic intervention in reducing gastric cancer risk and gastrointestinal disease burden, diagnostic accuracy remains paramount for effective treatment protocols. Current diagnostic methodologies encompass both invasive modalities (histopathological examination, microbiological culture, and rapid urease testing) and non-invasive approaches (serological antibody detection, 13C/14C-urea breath analysis, fecal antigen testing, and PCR-based DNA detection). Despite invasive techniques maintaining gold standard status in diagnostic algorithms, they present limitations including temporal inefficiency, cost considerations, and potential post-biopsy complications [29]. Conversely, non-invasive methodologies demonstrate susceptibility to pharmacological interference and inability to assess intragastric conditions [30]. The potential integration of

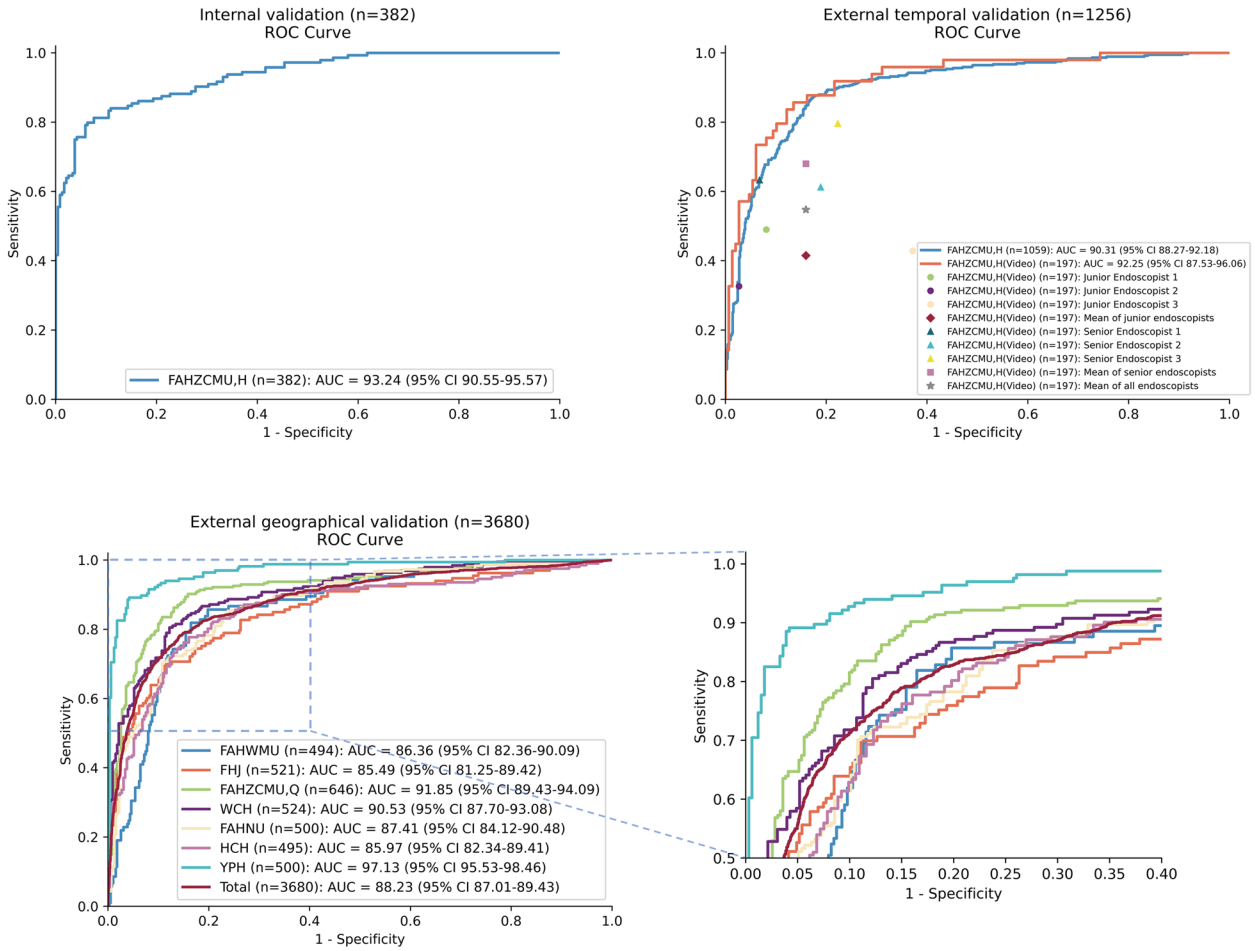

Fig. 3 Receiver operating characteristic (ROC) curves illustrate HOPE AI' ability to detect *H. pylori* infection. **A** ROC curve of internal cohort. **B** ROC curve of external temporal validation cohort. **C** ROC curve of external geographical validation cohort. ROC, receiver operating characteristic; FAHZCMU, H, The First Affiliated Hospital of Zhejiang Chinese Medical University, Hubin Campus; FAHWMU, The First Affiliated Hospital of Wenzhou Medical University; FHJ, The First Hospital of Jiaxing; FAHZCMU, Q, The First Affiliated Hospital of Zhejiang Chinese Medical University, Qiantang Campus; WCH, Wenzhou Central Hospital; FAHNU, The First Affiliated Hospital of Ningbo University; HCH, Huzhou Central Hospital, YPH, Yuyao People's Hospital of Zhejiang

endoscopic evaluation for concurrent assessment of gastric mucosal status and *H. pylori* infection status, without necessitating biopsy, represents a promising diagnostic paradigm. However, current endoscopic assessment of *H. pylori* infection status, while supported by qualitative mucosal characteristic documentation [4, 5, 31], remains constrained by inter-observer variability and subjective interpretation of endoscopic findings among clinicians.

Recent advancements in artificial intelligence applications within gastrointestinal endoscopy have revolutionized image recognition and processing capabilities, particularly in enhancing diagnostic accuracy for *Helicobacter pylori* infection [32, 33]. While conventional AI-assisted diagnostic systems predominantly utilize convolutional neural networks (CNNs) [10, 15, 34], these approaches face limitations in capturing comprehensive mucosal manifestations due to limited per-patient image sampling [35]. Furthermore, CNNs' reliance on image-level annotations for multi-feature *H. pylori* infection detection presents challenges in workflow efficiency and annotation reliability. To address these limitations and the inherent opacity of traditional AI models, our innovative system implements a transformer architecture incorporating self-attention mechanisms for case-level endoscopic image analysis. This integration with the MIL framework facilitates automated feature extraction and relationship mapping in latent space. The incorporation of LSTM networks enhances temporal pattern recognition through gated cell structures, enabling long-term state preservation.

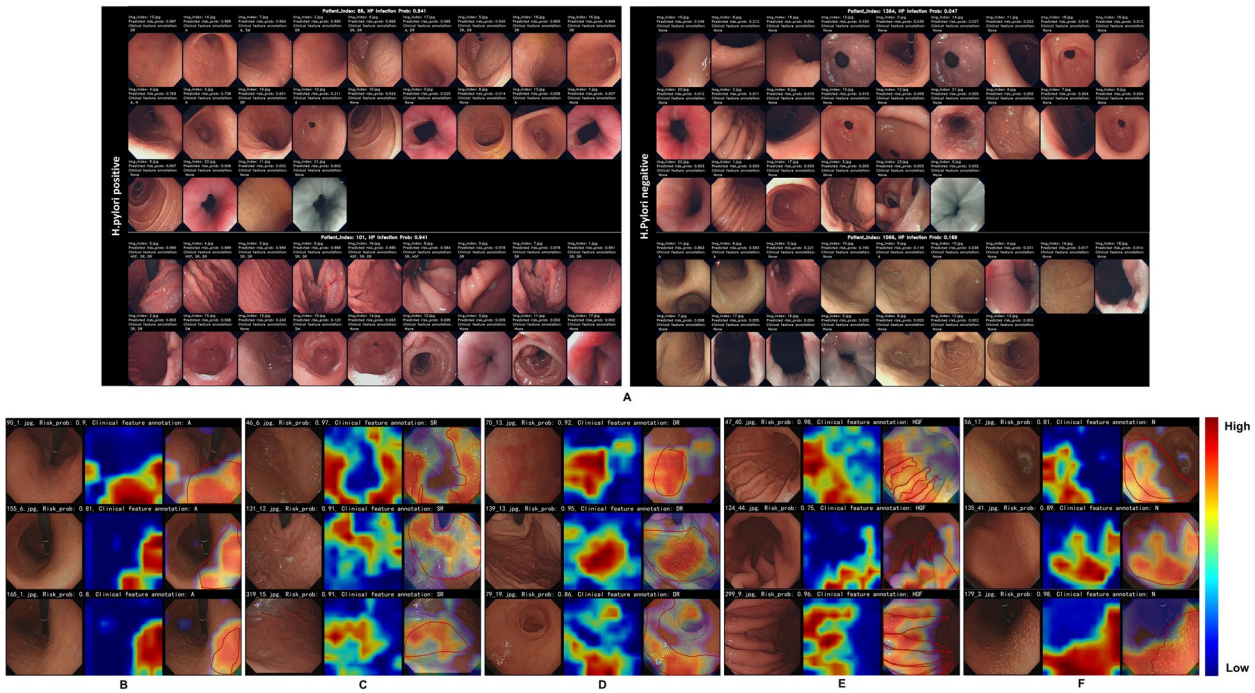

Fig. 4 Representative images generated from explainability and visualization analysis. **A** The model's prediction results with the endoscopists' assessments. All images are sorted according to the risk probabilities predicted by the model, and the judgments of endoscopy experts and the predicted probabilities of the model at the case level and image level are marked. **B**–**F** The visualization of the model's heatmaps at the pixel level by class activation mapping. The heatmap's color depth from blue (less focus) to red (more focus) positively contributed to its prediction as *H. pylori*-positive. The red outline represents the positive feature area marked by endoscopic experts. **A** atrophy, DR diffuse redness, SR spotty redness, SM sticky mucus, HGF hypertrophy of the gastric fold, N nodularity, RAC regular arrangement of collecting venules, FGP fundic gland polyp

Current *H. pylori* diagnostic methods typically analyze pre-selected, image-level features requiring manual or semi-automatic curation of relevant frames before annotation and model training [18, 19, 21]. This approach has significant drawbacks: heavy annotation requirements, inconsistent observer reliability, and compromised diagnostic accuracy from relying on limited image subsets rather than comprehensive assessment [10, 21]. In clinical practice, *H. pylori* diagnosis is fundamentally a case-level determination, while individual images within cases lack specific labels and vary greatly in quality and diagnostic relevance. Our MIL architecture addresses these challenges by treating each case as a "bag" of images, using attention mechanisms to assign importance scores to individual images before combining these weighted features into comprehensive case-level representations. This approach systematically identifies diagnostically important images and discriminative features driving final predictions. This methodology offers two key advantages: it eliminates time-consuming instance-level annotations and enhances robustness through attention-guided feature weighting that effectively filters out imaging noise (blur, defocus, overexposure, narrow-band imaging variations). This resilience makes our TransMIL + LSTM framework a practical and reliable computer-aided diagnostic system, particularly in clinical environments where obtaining perfect endoscopic data remains challenging.

The HOPE AI system demonstrated exceptional diagnostic capabilities in *H. pylori* detection, achieving sensitivity of 84.7% (95% CI 78.5–90.3%), specificity of 85.7% (81.4–91.2%), accuracy of 85.3% (81.68–88.74%), and AUC of 0.932 (0.906–0.956). The system's robust performance was maintained across diverse validation cohorts, encompassing both endoscopic imaging modalities, surpassing expert endoscopist assessment. While previous artificial intelligence applications in *H. pylori* endoscopic detection exist [36], they have been limited by methodological constraints including suboptimal golden standards [37], retrospective validation approaches [10], insufficient sample sizes [38], single-center designs [39], and restricted imaging modalities [18], impeding their clinical applicability and advancement. In alignment with Ghassemi et al.'s [40] emphasis on rigorous AI validation protocols, our investigation

incorporated prospective data acquisition from seven medical centers with dual external validation phases, establishing the system's robust generalizability across diverse clinical settings.

Notwithstanding the encouraging outcomes, several inherent limitations of the HOPE AI system warrant discussion. Primarily, the retrospective nature of data annotation in both training and internal validation cohorts may introduce selection bias, although the prospective external validation phase suggests minimal impact on system performance. Additionally, the absence of specific algorithmic approaches to address spatial correlation among frames from identical endoscopic video sequences could potentially introduce systematic bias. Nevertheless, HOPE AI demonstrated robust diagnostic accuracy across multiple clinical centers, indicating strong generalizability. Furthermore, the system's development and validation use predominantly Chinese patient data, which limits the possibility of generalizing the results and necessitates further investigation of its diagnostic performance in diverse ethnic populations.

Conclusions

In conclusion, our novel artificial intelligence–based endoscopic diagnostic system, HOPE AI, demonstrates superior diagnostic capabilities in *Helicobacter pylori* detection through multi-institutional validation utilizing diverse endoscopic imaging datasets. The system exhibits exceptional diagnostic accuracy with sensitivity metrics surpassing those of non-specialist practitioners. This robust performance positions HOPE AI as a promising computer-aided diagnostic tool for clinical endoscopists in *H. pylori* infection screening protocols. Future multicenter clinical trials and international validation studies are warranted to establish the system's generalizability and facilitate its global implementation in clinical gastroenterology practice.

Abbreviations

AI	Artificial intelligence
H. pylori	*Helicobacter pylori*
MIL	Multi-instance learning
LSTM	Long short-term memory
CI	Confidence interval
WHO	World Health Organization
IARC	International Agency for Research on Cancer
RAC	Regular arrangement of collecting venules
FAHZCMU, H	The First Affiliated Hospital of Zhejiang Chinese Medical University, Hubin Campus
WLE	White-light endoscopy
NBI	Narrow-band imaging
ViT	Vision transformer
PPV	Positive predictive value
NPV	Negative predictive value
ROC	Receiver operating characteristic
CNNs	Convolutional neural networks

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12916-025-04379-2.

Additional file 1: Supplementary Materials.

Additional file 2: Table S1–S3. Table S1: Information of endoscopic images in different sets. Table S2: Results of endoscopists for H. pylori diagnosis in video validation set. Table S3: Results of endoscopists vs. HOPE AI by sensitivity and specificity for H. pylori identification in video validation set (n=197).

Acknowledgements

We would like to thank all the endoscopists and medical engineers for their efforts. We also extend our thanks to Chao Chen, Yang Liu from The First Affiliated Hospital of Wenzhou Medical University, Yu Xiang from Huzhou Central Hospital, Honghui Chen, Yan Zhang from the First Affiliated Hospital of Ningbo University, Huang Su from Wenzhou Central Hospital, Li Chu from The First Hospital of Jiaxing, Jianbo Zhou, Jianzhong San from Yuyao People's Hospital of Zhejiang for the great assistance of data collection, and Jiajia Ying from the Cosmos Wisdom Biotech Co.Ltd for the great assistance of figure retouch.

Data transparency statement

Data and study materials are available from the corresponding author upon reasonable request.

Authors' contributions

Guarantor of article (BL, LZ, YH, JX, LH), study concept and study supervision (BL, LZ), data collection and/or data interpretation (JZ, TC, JL, FZ, XD, JP, XW, LH), methodology guidance (SL), data analysis (JX, ZZ), manuscript drafting (YH, JX). All authors have read and approved the manuscript.

Funding

This study was supported by "Pioneer" and "Leading Goose" R&D Program of Zhejiang (2023C03050).

Data availability

The datasets analyzed during the current study are not publicly available due to the identifiable patient information, but the de-identifiable data are available from the corresponding author on reasonable request. The code of HOPE AI can be found at [https://github.com/Scatteredrain/Helicobacter_Pylori-Infection-Detection] (https:/github.com/Scatteredrain/Helicobacter_Pylori-Infection-Detection) .

Declarations

Ethics approval and consent to participate

The study protocol was approved by the ethics committees of the First Affiliated Hospital of Zhejiang Chinese Medical University (No. 2024-KLS-489–01). Written informed consent was obtained from all participants or their legal representatives.

Consent for publication

All authors have agreed to the submission and publication of this study.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Gastroenterology, Zhejiang Provincial Hospital of Chinese Medicine), The First Affiliated Hospital of Zhejiang Chinese Medical University, 54 Youdian Road, Hangzhou 310006, China. [2]Key Laboratory of Digestive Pathophysiology of Zhejiang Province, The First Affiliated Hospital of Zhejiang Chinese Medical University, Hangzhou, China. [3]Alibaba Group, DAMO Academy, No.960-1 West Wenyi Road, Yuhang District, Hangzhou 310006, China. [4]Department of Gastroenterology and Hepatology, The First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China. [5]Department of Gastroenterology, Huzhou Central Hospital, Huzhou, China. [6]Department of Gastroenterology, Yuyao People's Hospital of Zhejiang, Yuyao, China. [7]Department

of Gastroenterology, The First Affiliated Hospital of Ningbo University, Ningbo, China. [8]Department of Gastroenterology, Wenzhou Central Hospital, Wenzhou, China. [9]Department of Gastroenterology, The First Hospital of Jiaxing, Jiaxing, China. [10]Center of Clinical Evaluation, The First Affiliated Hospital of Zhejiang Chinese Medical University, Hangzhou, China.

Received: 2 April 2025 Accepted: 2 September 2025
Published online: 08 October 2025

References

1. International Agency for Research on Cancer (IARC). Schistosomes, liver flukes and Helicobacter pylori. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Lyon, 7–14 June 1994. IARC Monogr Eval Carcinog Risks Hum. 1994;61:1–241.
2. Biological agents. Volume 100 B. A review of human carcinogens. IARC Monogr Eval Carcinog Risks Hum. 2012;100(Pt B):1–441. https://publications.iarc.who.int/Book-And-Report-Series/Iarc-Monographs-On-The-Identification-Of-Carcinogenic-Hazards-To-Humans/Review-Of-Human-Carcinogens-Package-Of-6-Volumes-A-B-C-D-E-F--2012.
3. Yan L, Chen Y, Chen F, Tao T, Hu Z, Wang J, et al. Effect of Helicobacter pylori eradication on gastric cancer prevention: updated report from a randomized controlled trial with 26.5 years of follow-up. Gastroenterology. 2022;163(1):154–62.e3. https://doi.org/10.1053/j.gastro.2022.03.039.
4. Kato M, Kamada T. Endoscopic findings for risk stratification of gastric cancer. In: Haruma K, Kato M, Inoue K, Murakami K, Kamada T. Kyoto Classification of Gastritis. 1st ed. Tokyo: Nihon Medical Center; 2017. p. 97–110.
5. Wang K, Zhao J, Jin H, Meng L, Fan Y, Zhou Y, et al. Establishment of modified Kyoto classification scoring model and its significance in the diagnosis of *Helicobacter pylori* current infection. Gastrointest Endosc. 2023;97:684–93.
6. Watanabe K, Nagata N, Shimbo T, Nakashima R, Furuhata E, Sakurai 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.
7. Chadebecq F, Lovat LB, Stoyanov D. Artificial intelligence and automation in endoscopy and surgery. Nat Rev Gastroenterol Hepatol. 2023;20:171–82.
8. Yuan XL, Liu W, Lin YX, Deng QY, Gao YP, Wan L, et al. Effect of an artificial intelligence-assisted system on endoscopic diagnosis of superficial oesophageal squamous cell carcinoma and precancerous lesions: a multicentre, tandem, double-blind, randomised controlled trial. Lancet Gastroenterol Hepatol. 2024;9:34–44.
9. Luo H, Xu G, Li C, He L, Luo L, Wang Z, 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.
10. Seo JY, Hong H, Ryu WS, Kim D, Chun J, Kwak MS. Development and validation of a convolutional neural network model for diagnosing *Helicobacter pylori* infections with endoscopic images: a multicenter study. Gastrointest Endosc. 2023;97:880-888.e2.
11. Jiang Y, Yan H, Cui J, Yang K, An Y. Artificial intelligence in endoscopy for predicting *Helicobacter pylori* infection: a systematic review and meta-analysis. Helicobacter. 2025;30:e70026.
12. Dilaghi E, Lahner E, Annibale B, Esposito G. Systematic review and meta-analysis: artificial intelligence for the diagnosis of gastric precancerous lesions and *Helicobacter pylori* infection. Dig Liver Dis. 2022;54:1630–8.
13. Bang CS, Lee JJ, Baik GH. Artificial intelligence for the prediction of *Helicobacter pylori* infection in endoscopic images: systematic review and meta-analysis of diagnostic test accuracy. J Med Internet Res. 2020;22:e21983.
14. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98:691–8.
15. Shichijo S, Nomura S, Aoyama K, Nishikawa Y, Miura M, Shinagawa T, et al. Application of convolutional neural networks in the diagnosis of *Helicobacter pylori* infection based on endoscopic images. EBioMedicine. 2017;25:106–11.
16. Wang W, Xie E, Li X, Fan D, Song K, Liang D, et al. PVT v2: improved baselines with pyramid vision transformer. Comput Vis Media. 2022;8:415–24.
17. Dong B, Wang W, Fan DP, Li J, Fu H, Shao L. Polyp-PVT: polyp segmentation with pyramid vision transformers. CAAI Artif Intell Res. 2023;2. https://doi.org/10.48550/arXiv.2108.06932.
18. Zhang M, Pan J, Lin J, Xu M, Zhang L, Shang R, et al. An explainable artificial intelligence system for diagnosing Helicobacter pylori infection under endoscopy: a case–control study. Therap Adv Gastroenterol. 2023;16:17562848231155024.
19. Li YD, Wang HG, Chen SS, Yu JP, Ruan RW, Jin CH, et al. Assessment of *Helicobacter pylori* infection by deep learning based on endoscopic videos in real time. Dig Liver Dis. 2023;55:649–54.
20. Zheng W, Zhang X, Kim JJ, Zhu X, Ye G, Ye B, 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. Lin CH, Hsu PI, Tseng CD, Chao PJ, Wu IT, Ghose S, et al. Application of artificial intelligence in endoscopic image analysis for the diagnosis of a gastric cancer pathogen-Helicobacter pylori infection. Sci Rep. 2023;13:13380.
22. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. In: Advances in neural information processing systems; 2019. p. 8024–35. https://doi.org/10.48550/arXiv.1912.01703.
23. Loshchilov I, Hutter F. SGDR: stochastic gradient descent with warm restarts. In Proc: 5th International Conference on Learning Representations; 2017. p.1769–84. https://doi.org/10.48550/arXiv.1608.03983.
24. Ilse M, Tomczak J, Welling M. Attention-based deep multiple instance learning. International conference on machine learning (PMLR). 2018.p. 2127–36. https://doi.org/10.48550/arXiv.1802.04712.
25. B Li, Y Li, KW Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021. https://doi.org/10.48550/arXiv.2011.08939.
26. Shao ZC, Biao H, Chen Y, Wang YF, Zhang J, Ji XY, et al. TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems. 2021. https://doi.org/10.48550/arXiv.2106.00908.
27. Ren S, Cai P, Liu Y, Wang T, Zhang Y, Li Q, et al. Prevalence of *Helicobacter pylori* infection in China: a systematic review and meta-analysis. J Gastroenterol Hepatol. 2022;37:464–70.
28. Hooi JKY, Lai WY, Ng WK, Suen MMY, Underwood FE, Tanyingoh D, et al. Global prevalence of *Helicobacter pylori* infection: systematic review and meta-analysis. Gastroenterology. 2017;153:420–9.
29. Patel SK, Pratap CB, Jain AK, Gulati AK, Nath G. Diagnosis of *Helicobacter pylori*: what should be the gold standard? World J Gastroenterol. 2014;20:12847–59.
30. Malfertheiner P, Megraud F, Rokkas T, Gisbert JP, Liou JM, Schulz C, et al. Management of *Helicobacter pylori* infection: the Maastricht VI/Florence consensus report. Gut. 2022. https://doi.org/10.1136/gutjnl-2022-327745.
31. Glover B, Teare J, Patel N. A systematic review of the role of non-magnified endoscopy for the assessment of *H. pylori* infection. Endosc Int Open. 2020;8:E105-14.
32. Lui TKL, Tsui VWM, Leung WK. Accuracy of artificial intelligence-assisted detection of upper GI lesions: a systematic review and meta-analysis. Gastrointest Endosc. 2020;92:821-830.e9.
33. Okagawa Y, Abe S, Yamada M. Artificial intelligence in endoscopy. Dig Dis Sci. 2022;67:1553–72.
34. Nakashima H, Kawahira H, Kawachi H, Sakaki N. 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.
35. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. WIREs Data Min Knowl Discov. 2019;9:e1312.
36. Parkash O, Lal A, Subash T, Sultan U, Tahir HN, Hoodbhoy Z, et al. Use of artificial intelligence for the detection of *Helicobacter pylori* infection from upper gastrointestinal endoscopy images: an updated systematic review and meta-analysis. Ann Gastroenterol. 2024;37:665–73.
37. Itoh T, Kawahira H, Nakashima H, Yata N. Deep learning analyzes *Helicobacter pylori* infection by upper gastrointestinal endoscopy images. Endosc Int Open. 2018;6:E139-44.

38. Yacob YM, Alquran H, Mustafa WA, Alsalatie M, Sakim HAM, Lola MSH. *H. pylori* related atrophic gastritis detection using enhanced convolution neural network (CNN) learner. Diagnostics (Basel). 2023;13:336.
39. Nakashima H, Kawahira H, Kawachi H, Sakaki N. Artificial intelligence diagnosis of *Helicobacter pylori* infection using blue laser imaging-bright and linked color imaging: a single-center prospective study. Ann Gastroenterol. 2018;31:462–8.
40. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3:e745–50.

Publisher's Note