J Gastroenterol (2023) 58:978–989
https://doi.org/10.1007/s00535-023-02025-3

The Japanese Society of Gastroenterology

ORIGINAL ARTICLE—ALIMENTARY TRACT

Exploring the challenge of early gastric cancer diagnostic AI system face in multiple centers and its potential solutions

Zehua Dong[1,2,3] · Xiao Tao[1,2,3] · Hongliu Du[1,2,3] · Junxiao Wang[1,2,3] · Li Huang[1,2,3] · Chiyi He[4] · Zhifeng Zhao[5] · Xinli Mao[6] · Yaowei Ai[7] · Beiping Zhang[8] · Mei Liu[9] · Hong Xu[10] · Zhenyu Jiang[11] · Yunwei Sun[12] · Xiuling Li[13] · Zhihong Liu[14] · Jinzhong Chen[15] · Ying Song[16] · Guowei Liu[17] · Chaijie Luo[1,2,3] · Yanxia Li[1,2,3] · Xiaoquan Zeng[1,2,3] · Jun Liu[1,2,3] · Yijie Zhu[1,2,3] · Lianlian Wu[1,2,3,18] · Honggang Yu[1,2,3,18]

Received: 2 May 2023 / Accepted: 10 July 2023 / Published online: 29 July 2023

Abstract

Background Artificial intelligence (AI) performed variously among test sets with different diversity due to sample selection bias, which can be stumbling block for AI applications. We previously tested AI named ENDOANGEL, diagnosing early gastric cancer (EGC) on single-center videos in man–machine competition. We aimed to re-test ENDOANGEL on multi-center videos to explore challenges applying AI in multiple centers, then upgrade ENDOANGEL and explore solutions to the challenge.

Methods ENDOANGEL was re-tested on multi-center videos retrospectively collected from 12 institutions and compared with performance in previously reported single-center videos. We then upgraded ENDOANGEL to ENDOANGEL-2022 with more training samples and novel algorithms and conducted competition between ENDOANGEL-2022 and endoscopists. ENDOANGEL-

Zehua Dong and Xiao Tao authors have contributed equally to this work.

✉ Lianlian Wu
wu_leanne@163.com

✉ Honggang Yu
yuhonggang@whu.edu.cn

[1] Renmin Hospital of Wuhan University, Wuhan, China

[2] Key Laboratory of Hubei Province for Digestive System Disease, Renmin Hospital of Wuhan University, Wuhan, China

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

[4] Department of Gastroenterology, Yijishan Hospital of Wannan Medical College, Wuhu 241001, Anhui, People's Republic of China

[5] Department of Digestive Endoscopy, The Fourth Hospital of China Medical University, Shenyang 110032, Liaoning Province, People's Republic of China

[6] Department of Gastroenterology, Taizhou Hospital of Zhejiang Province Affiliated to Wenzhou Medical University, Linhai, Zhejiang, China

[7] Department of Gastroenterology, The People's Hospital of China Three Gorges University, The First People's Hospital of Yichang, Yichang, China

[8] Department of Gastroenterology, The Second Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou, China

[9] Department of Gastroenterology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

[10] Department of Endoscopy, The First Hospital of Jilin University, Changchun, China

[11] Department of Gastroenterology, The Second Affiliated Hospital of Baotou Medical College, Baotou, Inner Mongolia, China

[12] Department of Gastroenterology, Ruijin Hospital, Shanghai Jiaotong University, Gubei Branch, Shanghai, People's Republic of China

[13] Department of Gastroenterology, School of Clinical Medicine, Henan Provincial People's Hospital, People's Hospital of Zhengzhou University, Henan University, Zhengzhou, Henan, China

[14] Department of Gastroenterology, Jilin City People's Hospital, Jilin, China

[15] Endoscopy Center, School of Medicine, The First Affiliated Hospital of Xiamen University, Xiamen University, Xiamen, China

2022 was then tested on single-center videos and compared with performance in multi-center videos; the two AI systems were also compared with each other and endoscopists.

Results Forty-six EGCs and 54 non-cancers were included in multi-center video cohort. On diagnosing EGCs, compared with single-center videos, ENDOANGEL showed stable sensitivity (97.83% vs. 100.00%) while sharply decreased specificity (61.11% vs. 82.54%); ENDOANGEL-2022 showed similar tendency while achieving significantly higher specificity (79.63%, $p < 0.01$) making fewer mistakes on typical lesions than ENDOANGEL. On detecting gastric neoplasms, both AI showed stable sensitivity while sharply decreased specificity. Nevertheless, both AI outperformed endoscopists in the two competitions.

Conclusions Great increase of false positives is a prominent challenge for applying EGC diagnostic AI in multiple centers due to high heterogeneity of negative cases. Optimizing AI by adding samples and using novel algorithms is promising to overcome this challenge.

Keywords Artificial intelligence · Early gastric cancer · Case diversity

Acronyms and abbreviations

AI	Artificial intelligence
GC	Gastric cancer
EGC	Early gastric cancer
LNM	Lymph node metastasis
SM	Submucosal
CNN	Conventional neural network
WLE	White light endoscopy
M-NBI	Magnifying narrow-band imaging
PPV	Positive predictive value
NPV	Negative predictive value

Introduction

Gastric cancer (GC) is the third leading cause of cancer-related mortality globally [1]. Most GCs are diagnosed at an advanced stage and have a much poorer prognosis than early gastric cancer (EGC) [2]. Early detection and curative treatment are key strategies for reducing GC-related mortality [3].

Artificial intelligence (AI) systems have achieved impressive ability in medical image analysis [4, 5]. However, the reported diagnostic performance of AI varied, which was partially caused by the heterogeneity and diversity among test sets due to sample selection bias [6, 7]. This leads to confusion in evaluating AI's performance and can be a stumbling block for the AI application in clinical practice. Besides, most previous studies tested AI systems or compared the performance of AI with endoscopists on still images [8, 9]; however, processing real-time data are crucial for AI, especially for endoscopic diagnosis, which enables real-time man–machine interactions [10, 11]. Therefore, exploring the difference between AI performance in different case diversity and going close to clinical application scenarios is essential.

In 2020, we conducted a man–machine competition between an AI system named ENDOANGEL and 46 endoscopists from 44 hospitals in China on detecting gastric neoplasms under white-light endoscopy (WLE) and diagnosing EGCs under magnifying endoscopy with narrow-band imaging (M-NBI) [12]. The test set of this competition was a single-center video cohort containing 100 videos consecutively collected from one institution. The source of these videos was relatively simple and may lack diversity. It is unknown how the ENDOANGEL will perform in a test set with greater diversity, whether there is a challenge, and what it is when adopting AI from a single center to multiple centers.

Therefore, in this study, we collected videos from 12 institutions with greater diversity and re-tested the ENDOANGEL in this multi-center video cohort. Then, we upgraded the ENDOANGEL to ENDOANGEL-2022 by enlarging training sets and adding novel training algorithms and held a second session of man–machine competition comparing the ENDOANGEL-2022 and endoscopists. Moreover, the ENDOANGEL was re-tested in a single-center video cohort and compared with the performance of ENDOANGEL to explore whether the influence of case diversity differs between the two AI systems. Both AI systems were compared with two groups of endoscopists in the two competitions. This is the first study exploring the difference between AI performance in different case diversity and the solutions of better adapting AI in real clinics.

[16] Department of Gastroenterology, Xi'an Gaoxin Hospital,
Xi'an 710032, Shaanxi Province, China

[17] Yi Xin Clinic, Changzhou, Jiangsu, China

[18] Department of Gastroenterology, Renmin Hospital of Wuhan
University, 99 Zhangzhidong Road,
Wuhan 430060, Hubei Province, China

Methods

Multi-center video collection and data pre-processing

In this study, a multi-center video cohort was retrospectively collected from 12 institutions in 8 provinces in China. Raw videos of pathological confirmed cancerous and noncancerous lesions observed both by WLE and M-NBI were enrolled. The detailed inclusion and exclusion criteria of videos and lesions were consistent with the previous study in 2020 and were listed in the Supplementary [12]. The pathological diagnosis and endoscopic videos were reviewed by three experienced gastroenterology pathologists with over ten years of experience in pathological diagnosis based on the WHO classification [13].

A research assistant reviewed all the videos and edited the raw videos into paired WLE and M-NBI video clips, which contain the best observation views of the lesions. All enrolled videos were stored in strict confidentiality by the research assistant until the formal competition. The single-center video cohort from the previous competition in 2020 was also set as a test set in this study.

Study design

This diagnostic study comprises three steps: (1) collect a cohort of multi-center videos, re-test the ENDOANGEL on multi-center videos; (2) upgrade the ENDOANGEL as ENDOANGEL-2022, conduct a man–machine competition between ENDOANGEL-2022 and endoscopists; and (3) re-test the ENDOANGEL-2022 in the previously collected single-center video cohort.

First, the ENDOANGEL was re-tested on the multi-center video cohort, and the test results were compared with that in the single-center video test, which was known from the previous study. An independent research assistant recorded the ENDOANGEL's answers in real time. Then, we upgraded the ENDOANGEL and renamed it ENDOANGEL-2022.

Second, we conducted a man–machine competition offline in Wuhan, China, on July 29, 2022. A group of endoscopists was invited and enrolled to participate in this competition and compared with ENDOANGEL-2022 on detecting gastric neoplasms and diagnosing EGCs using a multi-center video cohort. Here, the term "detecting" is also used interchangeably with "diagnosing". We employ this terminology to differentiate the diagnosis of gastric neoplasms from the diagnosis of cancer using M-NBI. All endoscopists reviewed videos using the same type of laptops and electronic answer sheets. The endoscopists and ENDOANGEL-2022 were required to answer whether the lesion was neoplasm or EGC when observing WLE and M-NBI video clips in real time.

Third, the ENDOANGEL-2022 was re-tested on the single-center video cohort from the first competition; its results were compared with that in the multi-center video cohort. Both the ENDOANGEL and ENDOANGEL-2022 were compared with endoscopists in the two competitions.

In this study, we defined an index named the error rate of each lesion/video, calculated as the number of endoscopists who misdiagnosed/the number of all endoscopists. The error rate of each lesion/video in each video cohort was calculated to assess the difficulty of the lesion. Furthermore, we set an error rate of 50% as the threshold for classifying the case in the video as typical (error rate $<$ 50%) or difficult (error rate $\geq$ 50%). The overall study design is shown in Fig. 1.

Participating endoscopists

The competition notice was posted to attract endoscopists nationwide to register for participation. We subsequently screened endoscopists who met the inclusion criteria similar to the previous competition described in the Supplementary. In addition, endoscopists from the hospitals that provided the videos were not allowed to participate to ensure the equity of the competition. The experience level of endoscopists was defined as junior, senior, and expert, with the experience of M-NBI $<$ 1 year, 1–3 years, $>$ 3 years. All enrolled endoscopists signed informed consent and completed a questionnaire for basic information collection.

Construction of ENDOANGEL-2022

The ENDOANGEL-2022 was upgraded from ENDOANGEL using 114,372 images, Mean-Teacher-based semi-supervised algorithms, and a novel image pre-processing and training method [14, 15]. The two AI systems shared the same functional modules of detecting gastric neoplasms under WLE (Convolutional neural network 1, CNN 1) and diagnosing EGC under M-NBI (CNN 2). The CNN model construction and test results were presented in the Supplementary.

Outcomes

The primary outcomes were the sensitivity of the ENDOANGEL in detecting neoplasms under WLE and diagnosing EGC under M-NBI videos. The secondary outcomes included the specificity, accuracy, positive predictive value (PPV), and negative predictive value (NPV) of ENDOANGEL for detecting neoplasms in WLE and

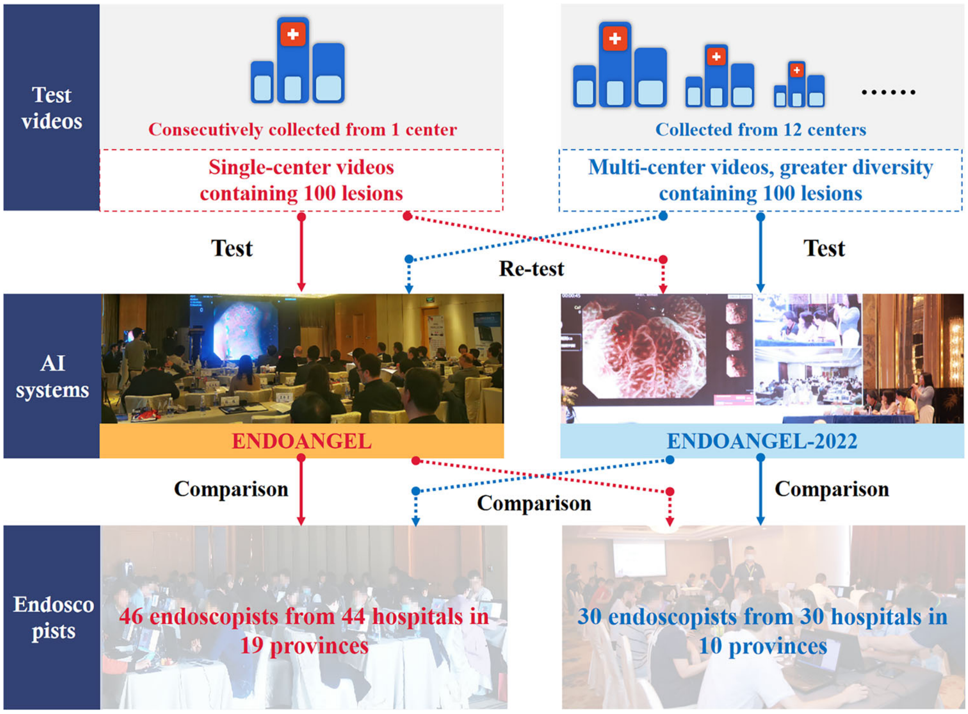
Test videos
Consecutively collected from 1 center
Single-center videos containing 100 lesions
Collected from 12 centers
Multi-center videos, greater diversity containing 100 lesions
Test
Re-test
Test
AI systems
ENDOANGEL
ENDOANGEL-2022
Comparison
Comparison
Comparison
Endoscopists
46 endoscopists from 44 hospitals in 19 provinces
30 endoscopists from 30 hospitals in 10 provinces

Fig. 1 Diagram of the study. In 2020, we held a man–machine competition using a single-center video cohort; the ENDOANGEL was tested and compared with 46 endoscopists. In 2022, we collected a multi-center video cohort and re-tested ENDOANGEL, then upgraded the ENDOANGEL to ENDOANGEL-2022. We then held another man–machine competition, comparing ENDOANGEL-2022 with 30 endoscopists. The ENDOANGEL-2022 was re-tested on single-center videos. Both AI were compared with the other cohort of endoscopists

diagnosing EGC in M-NBI; sensitivity, specificity, accuracy, PPV, and NPV of ENDOANGEL-2022 for detecting neoplasms in WLE and diagnosing EGC in M-NBI.

Ethics

This study was approved by the Ethics Committee of all the 12 institutions and was registered with trial number ChiCTR2200062192 in the WHO registry Network's Primary Registries. The institutional review boards exempted the informed consent for the retrospectively collected data.

Sample size

The sample size of test videos was calculated using the confidence intervals for one proportion method. Based on the previous study, we estimated that the sensitivity of detecting gastric neoplasm under WLE and diagnosing EGC under M-NBI was 85% for ENDOANGEL on a multi-center video cohort. With a type I error rate of 0.05 and a confidence interval width of 0.15, 99 videos were needed.

Statistical analysis

Performance metrics were compared between ENDOANGEL-2022, ENDOANGEL, and endoscopists in the two competitions using the Chi-square test, the McNemar test, and the Mann–Whitney *U* test. Performance metrics between different levels of endoscopists were compared using the Mann–Whitney *U* test. The inter-rater agreement among endoscopists was measured using Fleiss's Kappa. All p values are two sided with < 0.05 as significant.

Role of the funding source

The funder had no role in this study.

Results

The performance of ENDOANGEL in multi-center and single-center videos

Comparing the performance in single-center and multi-center videos, the ENDOANGEL showed stable sensitivity (87.81% vs. 91.07%, $p = 0.854$) on detecting gastric neoplasms under WLE, while the specificity decreased sharply (93.22% vs. 75.00%, $p < 0.05$). As for diagnosing EGCs under M-NBI, the ENDOANGEL also showed stable sensitivity (97.87% vs. 100.00%) and a sharply decreased specificity (61.11% vs. 82.54%, $p < 0.05$).

Characteristics of the patients, lesions, and endoscopists in this competition

In the present competition, 94 raw videos of 94 patients containing 100 lesions were retrospectively collected from 12 large referral centers in 8 provinces in China, including 46 EGCs and 54 non-cancers, 56 neoplasms, and 44 non-neoplasms. (Figure S1) The raw videos were edited into paired WLE and M-NBI video clips of 12.34 s [interquartile range (IQR), 10.20–14.35] and 59.67 s (IQR, 53.93–65.33). The characteristics of the patients and lesions are shown in Table S1. Forty-two endoscopists from 40 hospitals in 10 provinces in China registered for the competition. Twelve endoscopists declined to participate due to a time conflict. Ultimately, 30 endoscopists from 30 hospitals in 10 provinces in China came to Wuhan, China, and participated in the competition. (Figure S2) The baseline characteristics of endoscopists are presented in Table S2 and Table S3.

The characteristics of enrolled patients, lesions, and endoscopists in the previous competition were shown in the previous work [12].

Performance comparison of two AI systems and endoscopists

Detecting gastric neoplasms under WLE

When adapted from single center to multiple centers, the ENDOANGEL-2022 showed a similar tendency to the ENDOANGEL. The sensitivity of ENDOANGEL-2022 in the two video cohorts was stable (90.24% vs. 94.64, $p = 0.667$), while the specificity decreased (94.92% vs. 79.55%, $p < 0.05$) significantly. The decline in the specificity of ENDOANGEL-2022 was smaller.

On multi-center videos, the sensitivity (94.64%) and accuracy (88.00%) of ENDOANGEL-2022 were slightly higher than that of ENDOANGEL. And its accuracy (88.00% vs. 69.45%, $p < 0.001$), specificity (79.55% vs. 45.69%, $p < 0.001$), PPV (85.48% vs. 68.58%, $p < 0.001$), and NPV (92.11% vs. 80.50%, $p < 0.001$) were significantly higher than that of all endoscopists.

On single-center videos, the ENDOANGEL-2022 showed mildly higher accuracy, sensitivity, specificity, PPV, and NPV than ENDOANGEL. Moreover, ENDOANGEL-2022 outperformed endoscopists significantly in accuracy, sensitivity, specificity, PPV, and NPV. (Tables 1 and 2).

Diagnosing EGCs under M-NBI

When adapted from single center to multiple centers, the ENDOANGEL-2022 showed a similar tendency to the ENDOANGEL. The sensitivity of ENDOANGEL-2022 in the two video cohorts was stable (94.59% vs. 97.83%, $p = 0.848$), while the specificity decreased (90.48% vs. 79.63%, $p = 0.097$) but not significantly. The decline in the specificity of ENDOANGEL-2022 was smaller.

On multi-center videos, the accuracy (88.00% vs. 78.00%, $p = 0.013$) and specificity (79.63% vs. 61.11%, $p = 0.006$) of ENDOANGEL-2022 were significantly higher than that of ENDOANGEL. And its accuracy (88.00% vs. 74.53%, $p < 0.001$), sensitivity (97.83% vs. 88.70%, $p < 0.001$), specificity (79.63% vs. 62.47%, $p < 0.001$), PPV (80.36% vs. 68.79%, $p < 0.001$), and NPV (97.73% vs. 89.18%, $p < 0.001$) were significantly higher than that of all endoscopists.

On single-center videos, the ENDOANGEL-2022 showed mildly higher accuracy, specificity, and PPV than ENDOANGEL. And ENDOANGEL-2022 outperformed endoscopists significantly in accuracy, sensitivity, specificity, PPV, and NPV. (Tables 1 & 2).

Performance comparison of two cohorts of endoscopists

The two cohorts of endoscopists who participated in the single-center video test ($n = 46$) and the multi-center video test ($n = 30$) showed no significant difference in either characteristics or endoscopic experiences. (Table S3).

Cohort A showed higher inter-rater agreement than cohort B in detecting gastric neoplasms under WLE (Fleiss' kappa, 0.493 vs. 0.306). A similar finding was found in diagnosing EGC under M-NBI (Fleiss' kappa, 0.615 vs. 0.415). The sensitivity of cohort A for detecting gastric neoplasms under WLE was lower than cohort B [83.51% (95%CI, 81.23–85.79%) vs. 88.23% (95%CI, 83.60–92.86%), $p = 0.010$]. And the sensitivity of the two cohorts for diagnosing EGCs under M-NBI was comparable (87.13% vs. 88.70%, $p = 0.335$). Nevertheless, the specificity of the two cohorts for detecting gastric neoplasms and diagnosing EGCs showed much significant difference [72.33% vs. 45.69%, $p < 0.001$; 84.82% vs. 62.47%, $p < 0.001$]. (Fig. 2).

Error analysis of AI systems and endoscopists on single-center and multi-center videos

Detecting gastric neoplasms under WLE

As shown in Figs. 3 and 4, on multi-center and single-center videos, the two AI systems both missed LGD, which showed subtle changes in the mucosa, with a relatively high error rate of endoscopists. As for the false positives on non-neoplasms, the ENDOANGEL-2022 made fewer mistakes on typical lesions (error rate < 50%) than ENDOANGEL (one and three) in single-center videos. There were difficult lesions (error rate ≥ 50%) in both single-center and multi-center videos. AI systems made reasonable mistakes on lesions with more complicated changes, such as fold convergence or rough-surfaced mucosa, which were also difficult for endoscopists, as indicated by the error rate.

Table 1 Diagnostic performance of two AI systems and endoscopists in multi-center videos

Index	ENDOANGEL-2022 % (*n*/total)	ENDOANGEL % (*n*/total)	All endoscopists (n = 30) % (95%CI)	Experts (*n* = 9) % (95%CI)	Seniors (*n* = 12) % (95%CI)	Juniors (*n* = 9) % (95%CI)
Diagnosing neoplasm in WLE						
Sensitivity	94.64 (53/56)	91.07 (51/56)	88.23 (83.60–92.86)	90.48 (84.45–96.50)	86.02 (75.23–96.81)	88.69 (80.20–97.18)
Specificity	79.55 (35/44)	75.00(33/44)	45.69 (38.57–52.80)$^{***\wedge\wedge\wedge}$	45.96 (36.12–55.79)$^{***\wedge\wedge\wedge}$	50.20 (33.72–66.68)$^{**\wedge\wedge}$	39.90 (27.89–51.91)$^{***\wedge\wedge\wedge}$
Accuracy	88.00 (88/100)	84.00 (84/100)	69.45 (67.52–71.38)$^{***\wedge\wedge\wedge}$	70.89 (66.40–75.38)$^{***\wedge\wedge\wedge}$	70.09 (66.88–73.31)$^{***\wedge\wedge\wedge}$	67.22 (63.88–70.57)$^{***\wedge\wedge\wedge}$
PPV	85.48 (53/62)	82.26 (51/62)	68.58 (65.90–71.26)$^{***\wedge\wedge\wedge}$	68.36 (64.58–72.15)$^{***\wedge\wedge\wedge}$	71.14 (64.67–77.60)$^{**\wedge\wedge\wedge}$	65.67 (62.68–68.67)$^{***\wedge\wedge\wedge}$
NPV	92.11 (35/38)	86.84 (33/38)	80.50 (75.27–85.73)**	81.13 (70.14–92.13)	80.93 (71.46–90.40)	79.34 (68.26–90.42)
Diagnosing EGC in M-NBI						
Sensitivity	97.83 (45/46)	97.83 (45/46)	88.70 (83.93–93.46)$^{***\wedge\wedge\wedge}$	93.00(87.86–98.13)	85.69 (75.58–95.80)$^{*\wedge}$	88.41 (78.59–98.22)
Specificity	79.63 (43/54)	61.11 (33/54)**	62.47 (55.42–69.52)***	63.99 (51.60–76.38)	61.27 (47.30–75.23)**	62.55 (47.90–77.20)**
Accuracy	88.00 (88/100)	78.00 (78/100)*	74.53 (71.59–77.47)***	77.33 (71.86–82.81)***	72.50 (67.14–77.86)***	74.44 (68.43–80.46)***
PPV	80.36 (45/56)	68.18 (45/66)	68.79 (65.28–72.29)***	70.08 (63.19–76.97)	68.12 (61.18–75.06)**	68.38 (61.83–74.94)***
NPV	97.73 (43/44)	97.06 (33/34)	89.18 (85.85–92.52)$^{***\wedge\wedge\wedge}$	92.48 (87.87–97.10)**	86.82 (80.44–93.19)$^{**\wedge}$	89.04 (81.57–96.52)

PPV Positive predictive value; *NPV* negative predictive value

*Comparing with ENDOANGEL-2022, *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$

^Comparing with ENDOANGEL,$^{\wedge}$, $p < 0.05$; $^{\wedge\wedge}$, $p < 0.01$;$^{\wedge\wedge\wedge}$, $p < 0.001$

Table 2 Diagnostic performance of two AI systems and endoscopists in single-center videos

Index	ENDOANGEL-2022% (*n*/total)	ENDOANGEL % (*n*/total)	All endoscopists (n = 46) % (95%CI)	Experts (n = 8) % (95%CI)	Seniors (n = 19) % (95%CI)	Juniors (n = 19) % (95%CI)
Diagnosing neoplasm in WLE						
Sensitivity	90.24 (37/41)	87.81 (36/41)	83.51 (81.23, 85.79)***	81.40 (74.34, 88.46)	83.83 (80.2.9, 87.36)**	84.08 (80.26, 87.91)*
Specificity	94.92 (56/59)	93.22 (55/59)	72.33 (67.34, 77.31)$^{***\wedge\wedge}$	75.21 (60.72, 89.70)$^{***\wedge\wedge}$	77.34 (69.35, 85.33)$^{***\wedge}$	66.10 (58.61, 73.59)$^{***\wedge\wedge}$
Accuracy	93.00 (93/100)	91.00 (91/90)	76.91 (74.55, 79.27)$^{***\wedge\wedge}$	77.75 (71.47, 84.03)$^{***\wedge}$	80.00 (76.11, 83.89)$^{***\wedge}$	73.47 (70.14, 76.81)$^{***\wedge\wedge}$
PPV	92.50 (37/40)	90.00 (36/40)	70.56 (66.81, 74.30)$^{***\wedge}$	72.65 (60.85, 84.45)$^{***\wedge}$	74.80 (68.89, 80.70)$^{***\wedge}$	65.43 (60.12, 70.74)$^{***\wedge\wedge}$
NPV	93.33 (56/60)	91.67 (55/60)	86.76 (85.55, 87.98)***	86.09 (82.36, 89.83)***	87.58 (85.68, 89.48)***	86.23 (84.22, 88.24)***
Diagnosing EGC in M-NBI						
Sensitivity	94.59 (35/37)	100.00 (37/37)	87.13 (83.75, 90.51)**	83.78 (70.39, 97.18)$^{\wedge}$	89.33 (84.67, 94.00)*	86.34 (81.07, 91.62)
Specificity	90.48 (57/63)	82.54 (52/63)	84.82 (81.08, 88.55)**	91.47 (86.61, 96.33)	87.13 (81.06, 93.21)	79.70 (73.41, 85.99)
Accuracy	92.00 (92/100)	89.00 (89/100)	85.67 (83.69, 87.66)***	88.63 (85.34, 91.91)**	87.95 (84.67, 91.23)**	82.16 (79.22, 85.10)**
PPV	85.37 (35/41)	77.08 (37/48)	79.89 (76.30, 83.49)*	86.78 (80.38, 93.17)	82.94 (77.14, 88.74)	73.95 (68.44, 79.45)
NPV	96.61 (57/59)	100.00 (52/52)	92.59 (90.91, 94.26)***	91.40 (85.67, 97.13)	93.91 (91.49, 96.32)*	91.77 (89.01, 94.54)*

PPV Positive predictive value; *NPV* negative predictive value

*Comparing with ENDOANGEL-2022, *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$

^Comparing with ENDOANGEL, $^{\wedge}$, $p < 0.05$; $^{\wedge\wedge}$, $p < 0.01$; $^{\wedge\wedge\wedge}$, $p < 0.001$

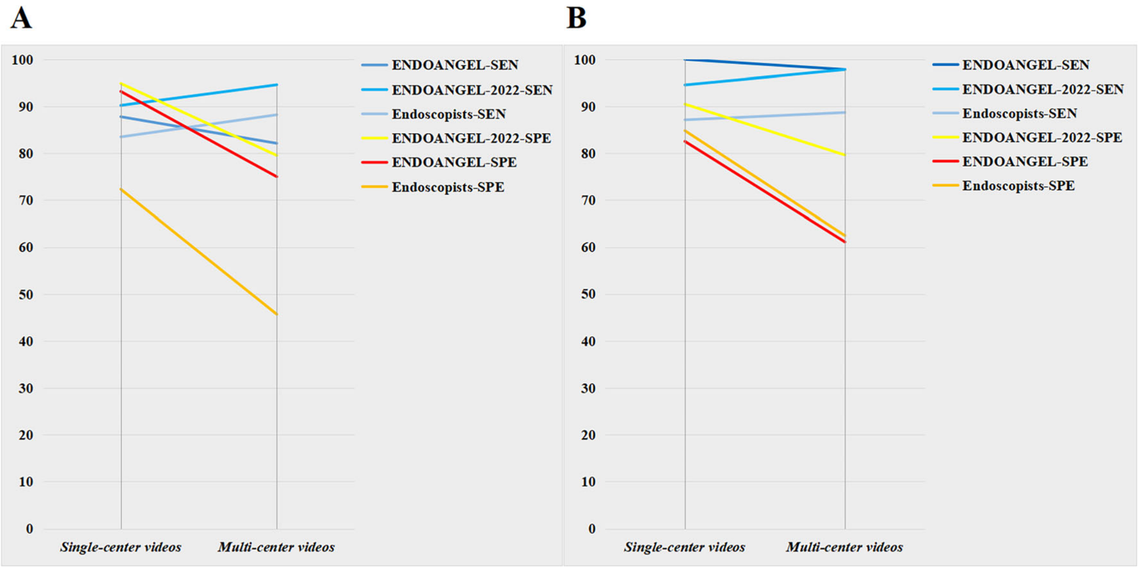

Fig. 2 The sensitivity and specificity of the two AI systems and endoscopists on single-center and multi-center videos. **A** Detecting gastric neoplasms under white-light endoscopy. **B** Diagnosing early gastric cancers under magnifying narrow-band imaging

Diagnosing EGC under M-NBI

As shown in Figs. 3 and 4, on multi-center and single-center videos, the two AI systems missed 4 EGCs. As for the non-cancers, the ENDOANGEL-2022 made fewer mistakes on typical lesions than that of ENDOANGEL in both single-center (two and six) and multi-center videos (three and eleven; $p < 0.05$, using McNemar Test). Both AI systems made reasonable mistakes on difficult lesions with irregular micro-vessels or micro-structures due to inflammation irritation, and lesions with more complicated changes such as irregular micro-vessels together with blood pigment or white coat, which were also difficult for endoscopists as indicated by the error rate.

Performance of AI and endoscopists stratified by *H. pylori* status

In multi-center videos, 67 patients had *H. pylori* infection test, 58.2% (39/67) were *H. pylori* negative and 41.8% (28/67) were *H. pylori* positive. On detecting neoplasms under WLE and diagnosing EGC under M-NBI videos, the endoscopists showed higher specificity (48.02% vs. 30.34%, $p < 0.01$; 66.33% vs. 43.94%, $p < 0.01$) and comparable sensitivity (86.48% vs. 90.22%, $p = 0.106$; 86.67% vs. 90.39%, $p = 0.056$) in *H. pylori*-negative cases compared with *H. pylori*-positive cases. No significant differences were observed in the performance of either ENDOANGEL-2022 or ENDOANGEL in patients with or without *H. pylori* infection. The diagnosis results are presented in Table S4.

Discussion

In this study, we re-tested an AI system named ENDOANGEL in a multi-center video cohort with greater case diversity and found that the specificity decreased sharply. We upgraded the AI, renamed it ENDOANGEL-2022, and conducted a second national man–machine competition. The ENDOANGEL-2022 was re-tested on the single-center video cohort and showed a similar performance tendency to the ENDOANGEL. Both AIs and endoscopists' performance fluctuated in test sets with different diversity. However, the ENDOANGEL-2022 showed better specificity and could be better adapted in multi-center videos.

AI-based image classification systems have been proven to have the potential to assist humans in lesion diagnosis [16, 17]. However, the heterogeneity and diversity among test sets may greatly affect the AI's performance, thus impairing its application in real clinic [18]. A test set sufficiently to cover real-world heterogeneity and diversity is crucial [19, 20]. As is well acknowledged, test samples derived from a single center may lack diversity and introduce selection bias; it is essential to fully test an AI system on a multi-center test cohort [21]. Moreover, it should be noted whether the test samples contain a large number of typical cases, which may cause overestimating of the AI's performance. For instance, Luo et al. developed an AI system diagnosing GC with a sensitivity of 94.0% and specificity of 96.1% [22]; however, advanced-stage GC lesions, which were obvious for recognizing, were included in the test set. Wu et al. developed an AI system that showed high sensitivity and specificity of 94.0% and

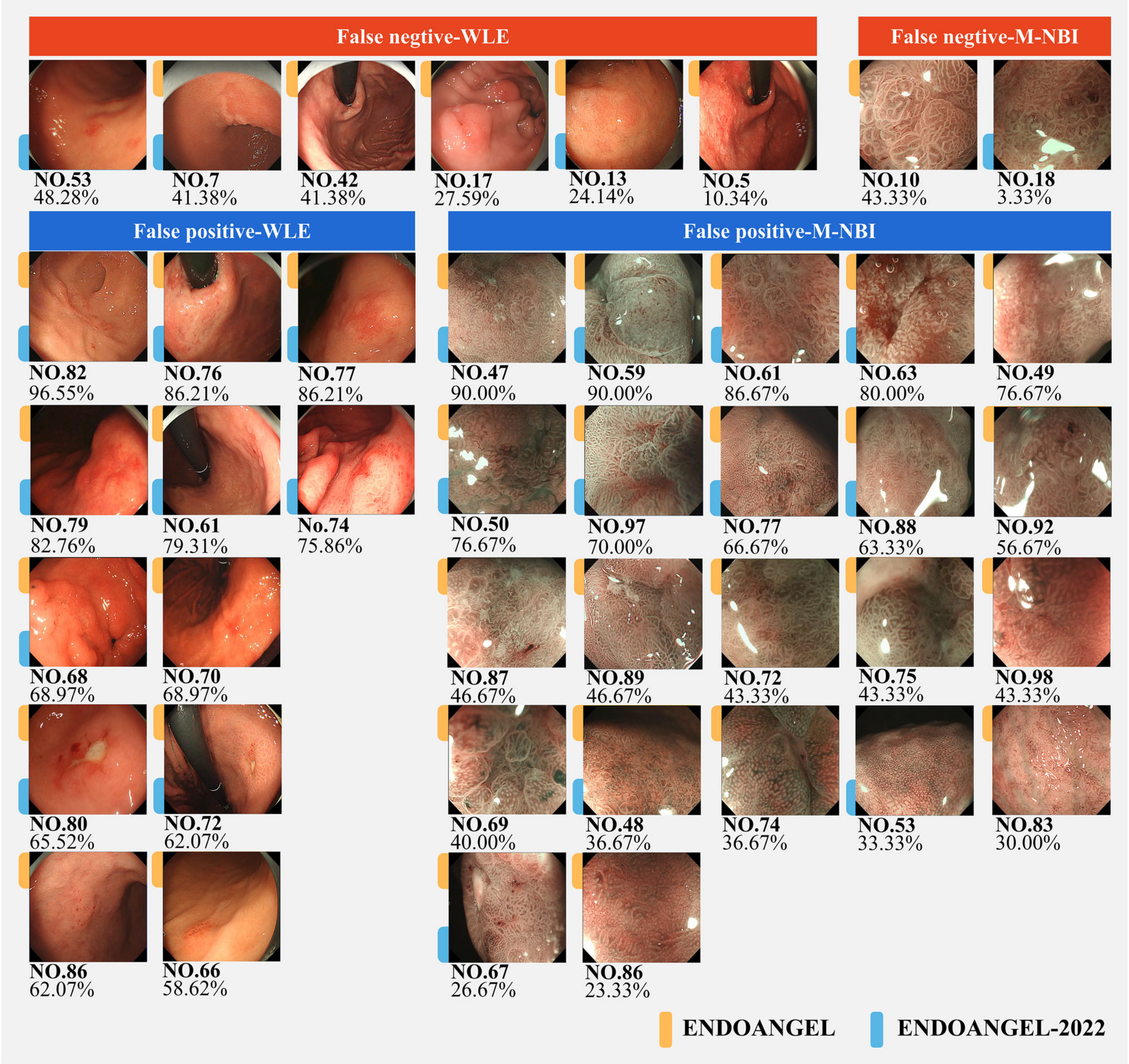

Fig. 3 Representative images of misdiagnosed lesions of two AI systems and endoscopists in multi-center videos. The error rate of endoscopists was presented below each case, and cases were set by the rank of error rate from high to low

91.0%, whereas only limited types of lesions (typical EGC, normal, superficial gastritis, and mild erosive gastritis mucosa) were enrolled in the test set [23]. In this study, we established a multi-center video cohort with good diversity containing videos of lesions needed to be observed under M-NBI to fully test the AI systems.

To test what we can do to solve the problem, we conducted technical explorations and upgraded the AI system. Inspired by the work for recognizing colorectal cancer in pathological images, we applied the Mean-Teacher-based semi-supervised algorithms for model construction [24]. Besides, according to another work, we applied a novel image pre-processing and training method by segmenting the original image into small tiles and outputting the final diagnosis of the whole image based on the prediction of each tile [14]. The results showed that the ENDOANGEL-2022 performed better than ENDOANGEL on detecting gastric neoplasms and diagnosing EGCs on both single-center and multi-center videos, especially on identifying non-cancers.

To explore what difference the multi-center video cohort introduced, we compared the performance between endoscopists in the two competitions and the performance of the ENDOANGEL-2022 when adapting from single center to

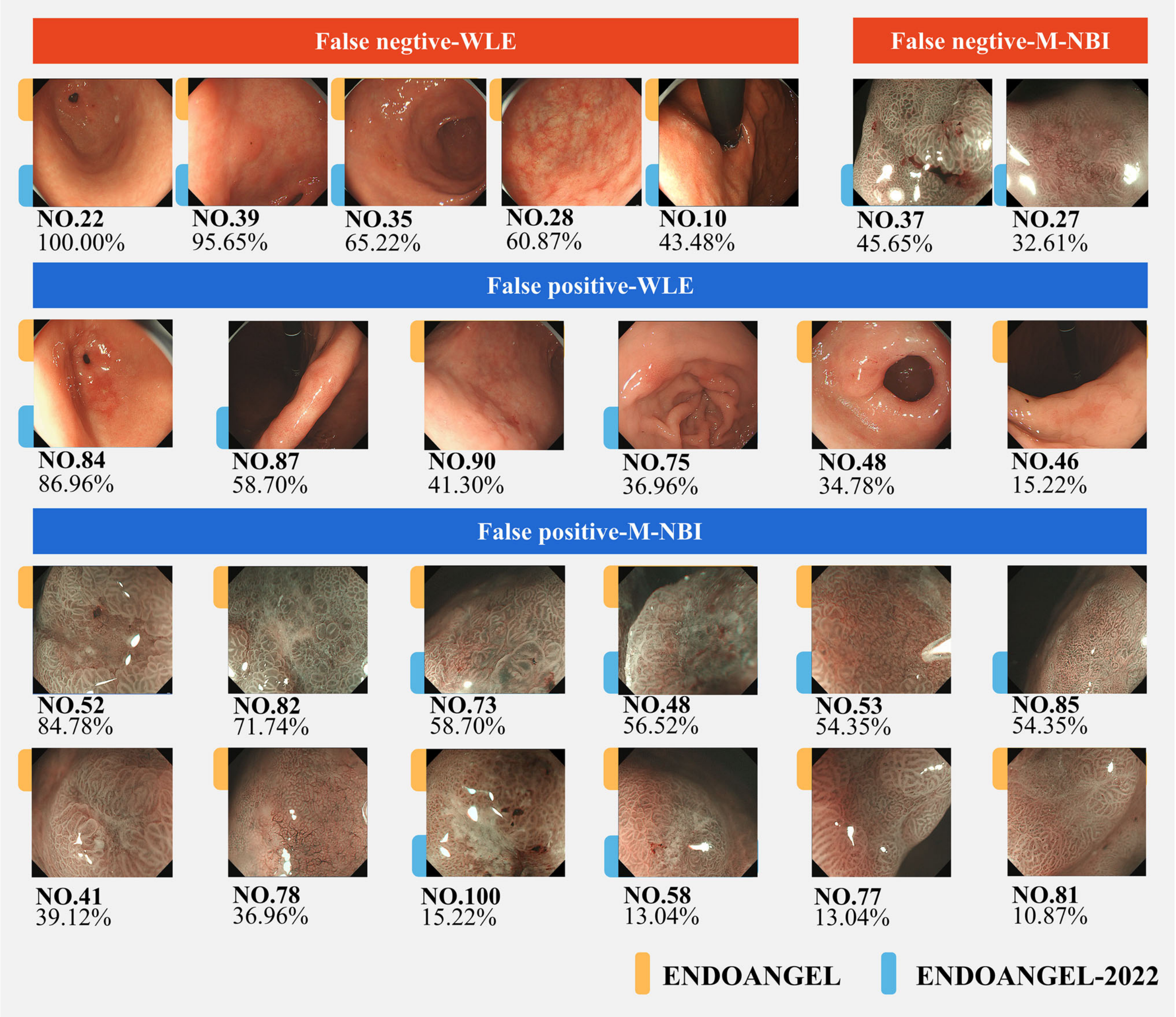

Fig. 4 Representative images of the misdiagnosed lesions of two AI systems and endoscopists in single-center videos. The error rate of endoscopists was presented below each case, and cases were set by the rank of error rate from high to low

multiple centers. As for the endoscopists, no significant differences were found between the two cohorts of endoscopists on indexes reflecting the characteristics and individual experience. However, the specificity of endoscopists in identifying negative cases differed significantly. In addition, compared with performance on single-center videos, the ENDOANGEL-2022 showed a similar tendency to the ENDOANGEL and endoscopists on multi-center videos, whose specificity decreased sharply while sensitivity remained stable. These findings indicated that the multi-center data led to greater heterogeneity of negative cases while less positive ones and caused more false positives of AI systems.

To explore what difference the ENDOANGEL-2022 showed, we compared the performance of the two AI systems in both single-center and multi-center video cohorts and analyzed the errors of the two AI systems. Both in the single-center and multi-center video cohort, the ENDOANGEL made mistakes in several typical cases due to interference, such as shadow, flexible angles when observing, and reflections. Errors on typical lesions may be annoying and damage the endoscopist’s trust in AI. However, ENDOANGEL-2022 showed higher specificity than ENDOANGEL; it misjudged fewer typical cases (error rate $< 50\%$) than ENDOANGEL both in single-center and multi-center videos (Fig. 4 and Fig. 3). Moreover, it remained the ability to distinguish noteworthy false positives sharing similar characteristics with neoplasms, such as severe inflammatory changes, which were hard for expert endoscopists to distinguish. These false positives are reasonable and may be of great value for reminding endoscopists of observation more seriously. It is promising for AI providers to pay more attention to dealing with the potential greater heterogeneity of negative cases and reducing distracting false positives in future development and promotion of AI.

The selection bias is a common problem in retrospective studies, while we have conducted strict strategies to minimize it. First, we have established detailed selection criteria for video enrollment and quality assurance, and the multi-center videos were sourced from diverse regional locations to avoid potential monogeneity and homogeneity issues that may arise from single-center videos. Second, all videos were provided by the study centers based on predefined inclusion/exclusion criteria, without any post hoc selection.

In this study, we explored the relationship between the ability of AI and endoscopists and patients' characteristics of *H. pylori* status. In our previous work exploring the relationship between the ability of AI and patients' characteristics of *H. pylori* infection in 749 patients [25], the AI achieved higher accuracy and specificity in patients without H pylori infection. However, the AI systems in this study did not demonstrate a statistically difference, which may be due to the limited sample size. But we did find that the endoscopists showed the same trend as the reported work. Combining the findings and the current results together, we hold the view that changes in the gastric mucosa resulting from *H. pylori* infection may impact the typical features of the lesion or introducing disturbing presences, leading to potential misdiagnosis.

Concerns about whether AI values when biopsies can be easily taken, and what are the benefits of AI when taking biopsy had been raised recently. First, we hold the view that endoscopic diagnosis before biopsy is imperative in clinical practice. On the one hand, performing biopsies on all detected abnormalities is time- and cost-consuming, and difficult to be implemented in routine clinic. Therefore, the guidelines recommend conducting targeted biopsies under M-NBI after endoscopic diagnosis [26]. On the other hand, it is important to choose a suitable area representing the mucosal changes among the detected abnormality to avoid underestimation on pathology. Moreover, it is widely acknowledged that the performance of endoscopists vary significantly [27]. Therefore, in our opinion, the benefits of AI include: (1) mitigating the gap between endoscopists' performance, reducing possible miss diagnosis and non-necessary biopsy. This was demonstrated by some previous studies, reporting that AI enhanced endoscopists' sensitivity and specificity, thereby minimizing instances of missed diagnoses and false positives [28, 29]. (2) improve the accuracy and efficiency of targeted biopsy.

There are some limitations in this study. First, this was a single-nation comparative study. Although we included videos and endoscopists from multiple centers and fully tested the performance of AI, further testing of the AI on international samples is required. Second, though we invited endoscopists from different levels of hospitals, few endoscopists came from primary hospitals. More endoscopists from primary hospitals in source-limited areas should be included in further studies. Third, we acknowledge that this was a preliminary study for assessing the AI generalization; further studies with larger scale and ensuring more cases from different institutions are needed. Fourth, though we explored the relationship of diagnostic performance and *H. pylori* infection, other characteristics such as the status of atrophy and the daily usage of drugs could not be analyzed to produce valid results due to data missing. Future studies exploring these issues are needed. Fifth, though we included videos from multiple centers and conduct man–machine comparison in this study, further assessment in real-time clinical setting is needed.

In conclusion, the increasing false-positive rate is a prominent challenge for applying EGC diagnostic AI from single to multiple centers due to high heterogeneity of negative cases. Optimizing AI by adding samples and using novel algorithms is promising to overcome this challenge. This study provided value and a basis for future studies applying AI in multiple centers.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00535-023-02025-3.

Acknowledgements College-enterprise Deepening Reform Project of Wuhan University (to Honggang Yu). Artificial Intelligence Application Demonstration Scenario Project Wuhan (grant no.2022YYCJ01) (to Honggang Yu). National Natural Science Foundation of China-Youth Science Fund (grant no.82202257) (To Lianlian Wu). Special projects for knowledge innovation of Wuhan (grant no.2022020801020482) (To Lianlian Wu).

Author contributions HGY and LLW made a big picture of the work and supervised the overall study. ZHD designed and did the experiments. ZHD and XT developed the system. CYH, ZFZ, XLM, YWA, BPZ, ML, HX, ZYJ, YWS, XLL, ZHL, JZC, YS, and GWL were involved in the data collection. ZHD and XT wrote the original draft. ZHD, HLD, CJL, LH, and XT analyzed the data. HGY and LLW revised the manuscript. JXW, XQZ, YXL, JL, and YJZ were involved in the format check and reviewing. HGY was responsible for the overall content as guarantor. All authors approved the final version of the report.

Funding College-enterprise Deepening Reform Project of Wuhan Universit, Artificial Intelligence Application Demonstration Scenario Project Wuhan, 2022YYCJ01, Hongang Yu, National Natural Science Foundation of China-Youth Science Fund, 82202257, Lianlian Wu, Special projects for knowledge innovation of Wuhan, 2022020801020482, Lianlian Wu.

Declarations

Conflict of interest The authors have nothing to declare.

References

1. Smyth EC, Nilsson M, Grabsch HI, et al. Gastric cancer. Lancet. 2020;396:635–48.
2. Katai H, Ishikawa T, Akazawa K, et al. Five-year survival analysis of surgically resected gastric cancer cases in Japan: a retrospective analysis of more than 100,000 patients from the nationwide registry of the Japanese gastric cancer association (2001–2007). Gastric Cancer. 2018;21:144–54.
3. Banks M, Graham D, Jansen M, et al. British society of gastroenterology guidelines on the diagnosis and management of patients at risk of gastric adenocarcinoma. Gut. 2019;68:1545–75.
4. Bi WL, Hosny A, Schabath MB, et al. Artificial intelligence in cancer imaging: clinical challenges and applications. CA Cancer J Clin. 2019;69:127–57.
5. Niazi MKK, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence. Lancet Oncol. 2019;20:e253–61.
6. Yu TF, He W, Gan CG, et al. Deep learning applied to two-dimensional color Doppler flow imaging ultrasound images significantly improves diagnostic performance in the classification of breast masses: a multicenter study. Chin Med J (Engl). 2021;134:415–24.
7. Haggenmuller S, Maron RC, Hekler A, et al. Skin cancer classification via convolutional neural networks: systematic review of studies involving human experts. Eur J Cancer. 2021;156:202–16.
8. Hirasawa T, Aoyama K, Tanimoto T, et al. Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images. Gastric Cancer. 2018;21:653–60.
9. Niikura R, Aoki T, Shichijo S, et al. Artificial intelligence versus expert endoscopists for diagnosis of gastric cancer in patients who have undergone upper gastrointestinal endoscopy. Endoscopy. 2022;54:780–4.
10. Johnson KB, Wei WQ, Weeraratne D, et al. Precision medicine, AI, and the future of personalized health care. Cts-Clin Transl Sci. 2021;14:86–93.
11. Chen M, Decary M. Artificial intelligence in healthcare: an essential guide for health leaders. Healthc Manage Forum. 2020;33:10–8.
12. Wu L, Wang J, He X, et al. Deep learning system compared with expert endoscopists in predicting early gastric cancer and its invasion depth and differentiation status (with videos). Gastrointest Endosc. 2022;95(92–104): e3.
13. Nagtegaal ID, Odze RD, Klimstra D, et al. The 2019 WHO classification of tumours of the digestive system. Histopathology. 2020;76:182–8.
14. Skrede OJ, De Raedt S, Kleppe A, et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. Lancet. 2020;395:350–60.
15. van Engelen JE, Hoos HH. A survey on semi-supervised learning. Mach Learn. 2020;109:373–440.
16. Maron RC, Haggenmuller S, von Kalle C, et al. Robustness of convolutional neural networks in recognition of pigmented skin lesions. Eur J Cancer. 2021;145:81–91.
17. Sharmat P, Hassan C. Artificial intelligence and deep learning for upper gastrointestinal neoplasia. Gastroenterology. 2022;162:1056–66.
18. Benkarim O, Paquola C, Park BY, et al. Population heterogeneity in clinical cohorts affects the predictive accuracy of brain imaging. PLoS Biol. 2022;20: e3001627.
19. Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med. 2020;26:1320–4.
20. Ryan M, Stahl BC. Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. J Inf Commun Ethics Soc. 2021;19:61–86.
21. de Hond AAH, Leeuwenberg AM, Hooft L, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. Npj Digital Med. 2022;5:2.
22. Luo H, Xu G, Li C, et al. Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: a multicentre, case-control, diagnostic study. Lancet Oncol. 2019;20:1645–54.
23. Wu L, Zhou W, Wan X, et al. A deep neural network improves endoscopic detection of early gastric cancer without blind spots. Endoscopy. 2019;51:522–31.
24. Yu G, Sun K, Xu C, et al. Accurate recognition of colorectal cancer with semi-supervised deep learning on pathological images. Nat Commun. 2021;12:6311.
25. 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. 2022;95:269-80.e6.
26. Chiu PWY, Uedo N, Singh R, et al. An Asian consensus on standards of diagnostic upper endoscopy for neoplasia. Gut. 2019;68:186–97.
27. Nakanishi H, Doyama H, Ishikawa H, et al. Evaluation of an e-learning system for diagnosis of gastric lesions using magnifying narrow-band imaging: a multicenter randomized controlled study. Endoscopy. 2017;49:957–67.
28. Tang D, Ni M, Zheng C, et al. A deep learning-based model improves diagnosis of early gastric cancer under narrow band imaging endoscopy. Surg Endosc. 2022;36:7800–10.
29. Yuan XL, Liu W, Liu Y, et al. Artificial intelligence for diagnosing microvessels of precancerous lesions and superficial esophageal squamous cell carcinomas: a multicenter study. Surg Endosc. 2022;36:8651–62.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.