Gastric Cancer (2023) 26:275–285
https://doi.org/10.1007/s10120-022-01358-x

ORIGINAL ARTICLE

A deep-learning based system using multi-modal data for diagnosing gastric neoplasms in real-time (with video)

Hongliu Du[1,2,3] · Zehua Dong[1,2,3] · Lianlian Wu[1,2,3] · Yanxia Li[1,2,3] · Jun Liu[1,2] · Chaijie Luo[1,2,3] · Xiaoquan Zeng[1,2,3] · Yunchao Deng[1,2,3] · Du Cheng[1,2,3] · Wenxiu Diao[1,2,3] · Yijie Zhu[1,2,3] · Xiao Tao[1,2,3] · Junxiao Wang[1,2,3] · Chenxia Zhang[1,2,3] · Honggang Yu[1,2,3]

Received: 24 October 2022 / Accepted: 25 November 2022 / Published online: 15 December 2022

Abstract

Background White light (WL) and weak-magnifying (WM) endoscopy are both important methods for diagnosing gastric neoplasms. This study constructed a deep-learning system named ENDOANGEL-MM (multi-modal) aimed at real-time diagnosing gastric neoplasms using WL and WM data.

Methods WL and WM images of a same lesion were combined into image-pairs. A total of 4201 images, 7436 image-pairs, and 162 videos were used for model construction and validation. Models 1–5 including two single-modal models (WL, WM) and three multi-modal models (data fusion on task-level, feature-level, and input-level) were constructed. The models were tested on three levels including images, videos, and prospective patients. The best model was selected for constructing ENDOANGEL-MM. We compared the performance between the models and endoscopists and conducted a diagnostic study to explore the ENDOANGEL-MM's assistance ability.

Results Model 4 (ENDOANGEL-MM) showed the best performance among five models. Model 2 performed better in single-modal models. The accuracy of ENDOANGEL-MM was higher than that of Model 2 in still images, real-time videos, and prospective patients. (86.54 vs 78.85%, $P=0.134$; 90.00 vs 85.00%, $P=0.179$; 93.55 vs 70.97%, $P<0.001$). Model 2 and ENDOANGEL-MM outperformed endoscopists on WM data (85.00 vs 71.67%, $P=0.002$) and multi-modal data (90.00 vs 76.17%, $P=0.002$), significantly. With the assistance of ENDOANGEL-MM, the accuracy of non-experts improved significantly (85.75 vs 70.75%, $P=0.020$), and performed no significant difference from experts (85.75 vs 89.00%, $P=0.159$).

Conclusions The multi-modal model constructed by feature-level fusion showed the best performance. ENDOANGEL-MM identified gastric neoplasms with good accuracy and has a potential role in real-clinic.

Keywords Gastric neoplasms · Multi-modal fusion · Weak magnification endoscopy · Deep learning

Introduction

Gastric cancer (GC) ranks fifth in the incidence of malignant tumors worldwide, with 770,000 related deaths in 2020 globally [1]. GC is symptomless at an early stage and most individuals are not diagnosed until the disease has progressed [2]. The five-year survival rate was less than 40% for advanced GC, but greater than 90% for GC detected at an early stage [3]. Early diagnosis is essential for patients' welfare and contributes to mitigating the economic burden on the health care system [4].

Digestive endoscopy is the first-line method to detect early gastric cancer (EGC) [5–7]. White light endoscopy is most commonly used to identify general features of suspected lesions, whereas it is difficult to distinguish the subtle

Hongliu Du, Zehua Dong and Lianlian Wu have contributed equally to this work.

✉ Honggang Yu
yuhonggang1968@163.com

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

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

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

changes of the mucosa and its sensitivity and specificity of EGC are 0.48 (95% CI 0.39–0.57) and 0.67 (0.62–0.71) [8, 9]. Dye-based chromoendoscopy could be effective in detecting lesions and is also recommended by the guideline to recognize high-risk lesions, but the technique is time-consuming and its uptake is limited [10]. Image-enhanced endoscopy (IEE) uses narrow-band spectrum or blue laser imaging to enhance the structural features of blood vessels and glands of gastric mucosa, which improves diagnostic accuracy [11]. Magnifying-IEE (M-IEE) has satisfied diagnostic ability for gastric neoplasms, however, its high cost of equipment and strict training requirements for endoscopists limit its popularity [12]. Weak magnifying-IEE (WM-IEE) achieves better performance than WL, with wide utility and relatively lower cost than M-IEE, provides a significant option for diagnosis of high-risk lesions such as gastric neoplasms [13].

Deep learning (DL) had showed great potential in medical image analysis and was reported to effectively assist physicians in disease diagnosis and treatment. In recent years, various studies had used deep learning to identify EGC and gastric neoplasms, but most of them focused on WL or magnifying endoscopy (ME)-IEE, while little concentrated on WM-IEE [14–16]. In addition, previous studies focused on single-modal data rather than multi-modal data, which may lead to incomplete identification of lesions and neglect of modality-specific information among different imaging modalities [17]. In actual clinical practice, guidelines also recommend the use of multi-modal light sources with chromoendoscopy and white light endoscopy instead of single light source [18].

In the present study, we developed three DL models using multi-modal fusion methods (WL incorporated with WM) for diagnosing gastric neoplasms and compared them with single-modal models (WL only and WM only). The five DL models were tested on three test levels including image, video, and prospective patients. The best model was selected to construct ENDOANGEL-MM (Multi-modal). We compared the performance between the models and endoscopists and conducted a diagnostic study to explore the ENDOANGEL-MM's effectiveness on improving the endoscopists' ability (Fig. 1). To our best knowledge, this is the first study to develop a deep-learning based system for diagnosing gastric neoplasms using WM and WL integrated data and to explore the optimal method of multi-modal data fusion.

Methods

Four datasets were used for training, validation, and testing the models, including training and validating set (Dataset 1), image test set (Dataset 2), video test set (Dataset 3), and prospective test set (Dataset 4). The detailed information of four dataset were presented in the supplementary materials.

Inclusion criteria: the lesions were viewed at WL and WM mode. Exclusion criteria: the lesions were hard to evaluate because of poor-quality views, resulting from active bleeding, thick white coats, blurs, defocus, mucus, and so on.

Two senior endoscopists were involved in labeling images, selecting and editing videos, both of whom had an experience of EGD over 5 years.

The equipment used in this study included standard gastroscopes [(EG-L590ZW; Fujifilm, Tokyo, Japan), (GIF-HQ290, GIF-H260Z, GIF-H290Z; Olympus Medical Systems, Tokyo, Japan)] and video systems [(ELUXEO 7000, LASEREO7000 and VP-4450HD; Fujifilm, Tokyo, Japan), (EVIS LUCERA CV-260/CLV-260 and EVIS LUCERA ELITE CV-290/CLV-290SL; Olympus Medical Systems, Tokyo, Japan)].

Development of five models

WM mode, such as the Near Focus mode of Olympus, is under a magnification of about 45x. Figure 2 shows representative images captured under WM mode.

To test and compare the diagnostic performance of WL only, WM only, WL and WM multi-modal fusion on task-level, feature-level, and input-level, Model 1, Model 2, Model 3, Model 4, and Model 5 were constructed, respectively (Fig. 3).

Model 1 and Model 2 were constructed to independently diagnose gastric neoplasms under WL and WM mode. 1176 WL images and 2273 WM images from dataset 1 were used to train and validate Model 1 and Model 2, respectively.

Model 3 (Multi-modal fusion on task-level): Here we regarded the diagnosis process of WL and WM model as "tasks". This data-fusion method followed the diagnosis logic of endoscopists in clinical practice. We combine WL and WM models in a tandem way. One lesion judged as neoplasm by Model 1 will be further judged by Model 2, and the result of Model 2 will be taken as the final answer of this lesion. If the lesion was regarded as non-neoplasm by Model 1, the lesion would not be sent to Model 2 and the answer of Model 1 will be taken as the final answer.

Model 4 (Multi-modal fusion on feature-level): The workflow of CNN (conventional neural network)-based AI models can be disassembled as the following steps: data input, feature extraction, classification and prediction output. Here, the WL and WM images of a same lesion were combined into image-pairs, and the data-fusion process was achieved at the feature-extraction step: the WL image and WM image were separately inputted in two independent CNN models for feature extraction; then, the features extracted from the

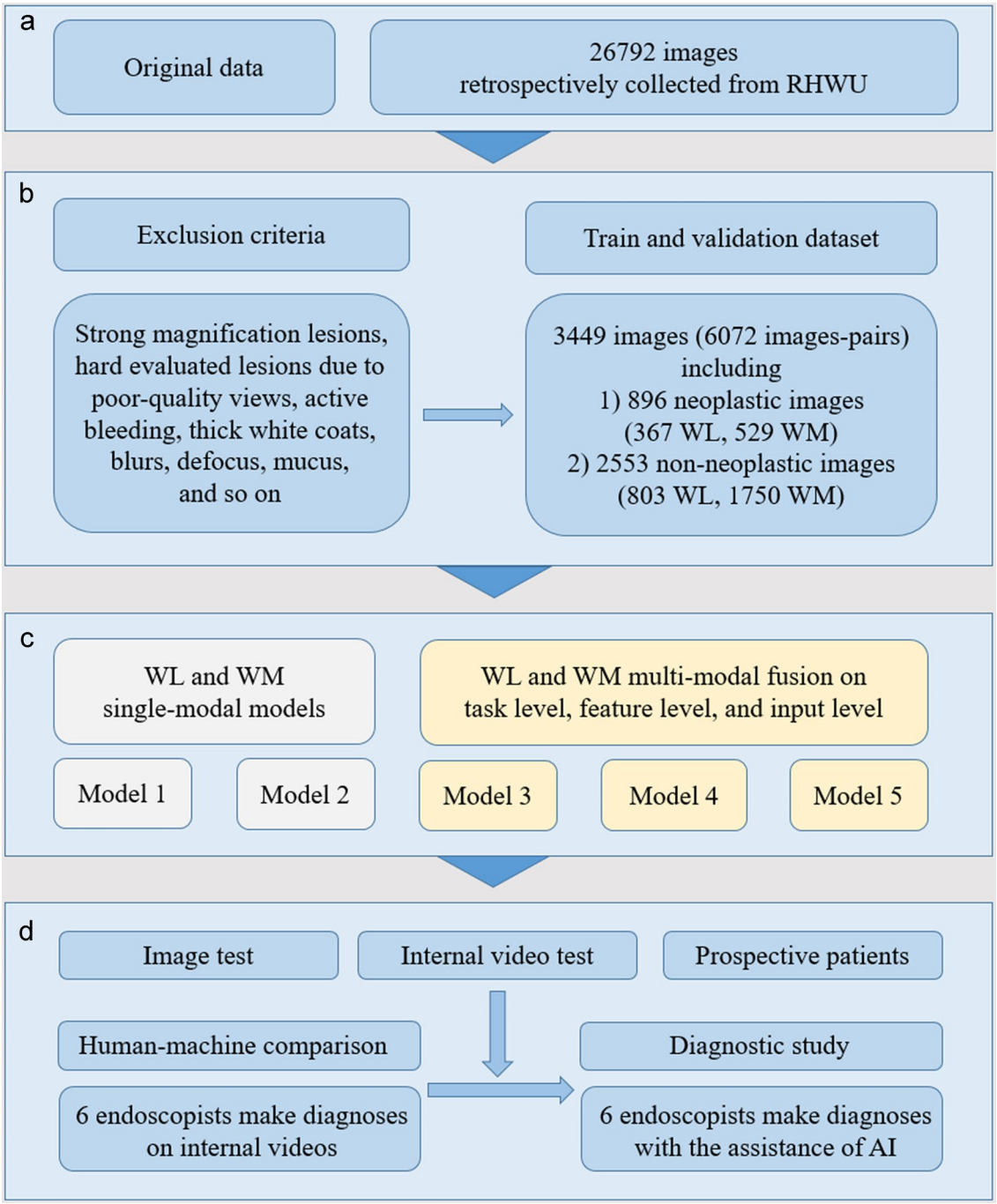

Fig. 1 Flow chart of Endoangel-MM (multi-modal). **a** Original data. **b** Data preprocessing. **c** Model construction. Two single-modal models and three multi-modal models were constructed. Model 1, white light (WL) single-modal. Model 2, weak magnification (WM) single-modal. Model 3, multi-modal fusion on task level. Model 4, multi-modal fusion on feature level. Model 5, multi-modal fusion on input level. **d** Tests of models. The five models were tested on three data levels including image, image pairs, and lesions, and three test levels including image, video, and prospective patients. A man–machine comparison and a case-reading study were conducted. The best model was selected for constructing Endoangel-MM and used in diagnostic case-reading study

two types of images were combined and used for further learning and classification.

Model 5 (Multi-modal fusion on input-level): As aforementioned, the basic workflow of CNN begins with data input and ends with prediction output. Here the data-fusion process was achieved at the data-input step. Similarly, the WL and WM images of a same lesion were combined into image-pairs. However, they were integrated and inputted to

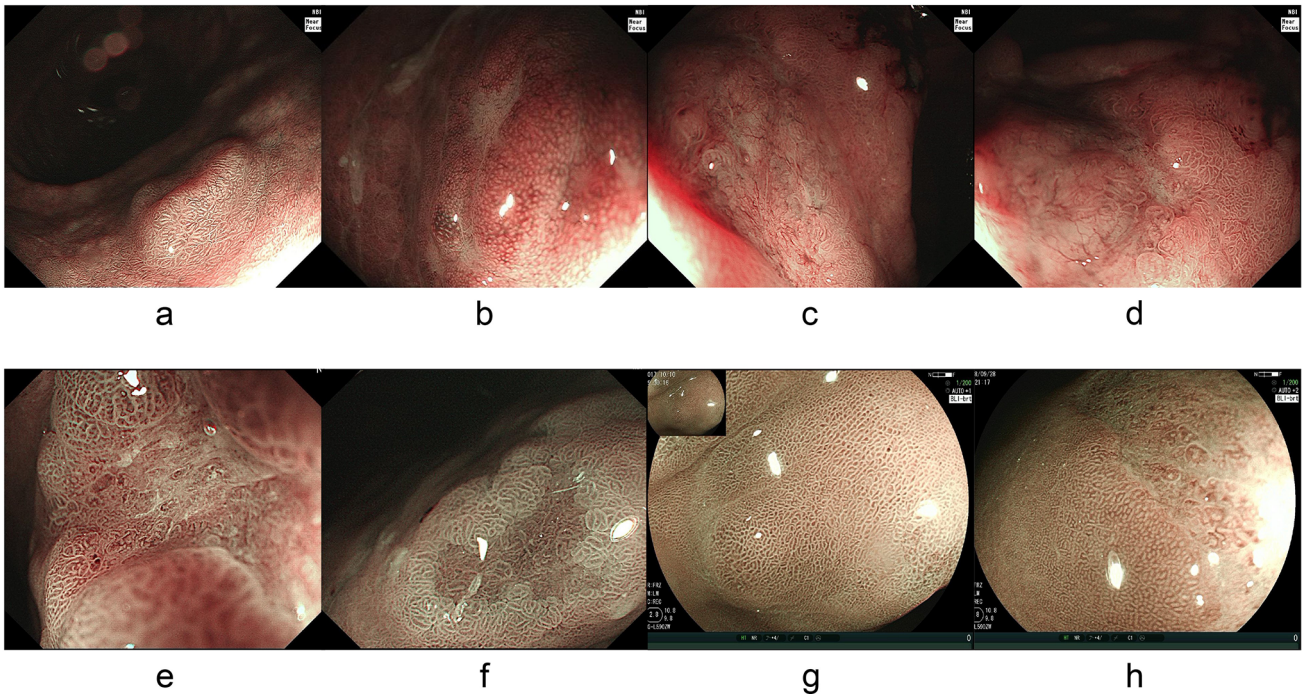

Fig. 2 Representative images of weak magnification mode. To increase the applicability, we incorporated the image data of magnifying narrow band imaging (M-NBI) and magnifying blue laser imaging (M-BLI) with the same magnification range as NF mode (45 × optical magnification) to construct the weak magnification dataset. **a**–**d** represent weak magnification images captured by near-focus mode. **e**, **f** represent weak magnification images viewed by M-NBI. **g**, **h** represent weak magnification images viewed by M-BLI

the CNN model as a whole (spliced image pair), and the feature-extraction process was done based on the spliced image pair.

The details of model training and constructing methods were presented in the supplementary materials.

Image test

Dataset 2 was used to test the performance of Model 1–5 in diagnosing gastric neoplasms in still images.

Model 1 makes a diagnosis in WL mode while Model 2 is in WM mode; both give a diagnosis at the image level and the lesion level. For Model 1 & 2, each lesion would be judged as neoplastic when at least one image is judged to be neoplastic; otherwise, the lesion will be judged as nonneoplastic.

Model 3, Model 4 and Model 5 judge the neoplasms based on image-pairs of lesions and thus could diagnose at both image-pair level and lesion level.

Video test

Dataset 3 and Dataset 4 were used to test the performance of Models 1–5 in diagnosing gastric neoplasms in raw videos. Each video clip is captured as a set of 7 frames per second, and each Model gives a diagnosis at the level of the lesions, with the same diagnostic logic as mentioned. During the WL clip, Model 4 and Model 5 won't make diagnosis but only store WL images. When the video clip shifts to the WM clip, Model 4 and Model 5 will store WM images and those WM images were made into imagepairs with stored WL images. Then the diagnostic result of each image-pair and each lesion will be output.

Comparing the performance of different models and endoscopists

Six endoscopists with at least 2 years of EGD experience (including 2 experts with more than 5 years of EGD experience, 2 seniors with 3~5 years of EGD experience, and 2 juniors with 1~2 years of EGD experience) were invited to participate in a man–machine competition using dataset 3. In the first round of testing (WM single-modal test), only WM videos were provided to endoscopists. The diagnostic performance of endoscopists was compared to WM single-modal model (Model 2). The second round of testing (multi-modal test) was performed 3 weeks later, and both WM and WL videos were provided to endoscopists. The diagnostic performance of endoscopists was compared to the best performing model among multi-modal models (Model 4).

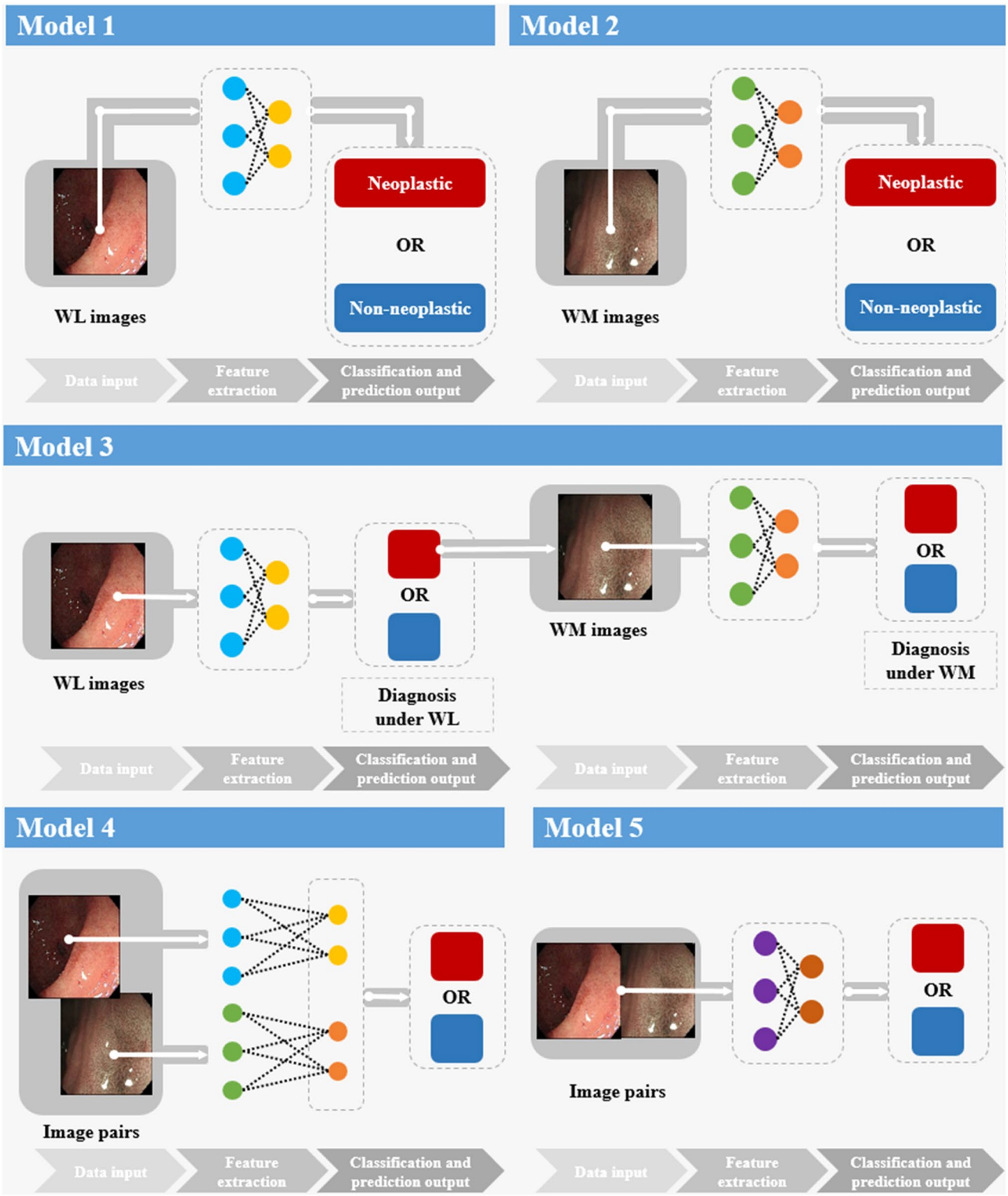

Fig. 3 Construction of Models. Model 1 diagnoses gastric neoplasms under white light (WL). Model 2 diagnoses gastric neoplasms under weak magnification (WM). Model 3 is an integration of Model 1 and Model 2 on task level, only the images/lesions diagnosed as neoplasms by Model 1 will be sent to Model 2 for further judgment. Model 4 is an integration of WL data and WM data on feature level, which makes decisions on image-pairs and lesions' level. The WL image and WM image were separately inputted in two independent CNN models for feature extraction. Model 5 is a fusion of WL data and WM data on input level. The WL and WM images of a same lesion were combined into image-pairs and inputted to the CNN model as a whole (spliced image pair), and the feature-extraction process was done based on the spliced image pair

Comparing the performance with and without ENDOANGEL-MM's assistance

A diagnostic study was conducted. Another 3 weeks after the second round of man–machine comparison, the diagnostic results of the ENDOANGEL-MM. Another 3 weeks after the second round of man-machine comparison, the best model among Model 1-5 will be selected as ENDOANEL-MM, and will be used for assistance in the following dignostic study. The endoscopists were allowed to read WL videos, WM videos, and the judgment of ENDOANGEL-MM, finally making a comprehensive diagnosis. Then the diagnostic performance of endoscopists in the second round (without ENDOANGEL-MM assistance) and third round (with ENDOANGEL-MM assistance) was compared.

Subgroup analysis

As advanced gastric cancer is easier to detect because of its typical characteristics, we conducted a sub-analysis stratified by excluding advanced gastric cancer in dataset 2, 3, and 4.

Outcomes

The objectives of this study were to evaluate the capabilities of ENDOANGEL-MM to diagnose gastric neoplastic lesions, the assistance of AI in improving the diagnostic performance of endoscopists, and the performance of multi-modal model and single-modal model.

To evaluate the capabilities of ENDOANGEL-MM and endoscopists, the accuracy, sensitivity, and specificity for diagnosing gastric neoplasms were calculated as follows: Accuracy = true predictions/total number of cases, sensitivity = true positive/positive, specificity = true negative/negative.

Ethics

This study was approved by the Ethics Committee of RHWU. Informed consent was exempted by the institutional review board for the retrospective data.

Statistical analysis

As for the prospective video test, the accuracy of ENDOANGEL-MM was estimated at 90%. The sample size was calculated as 62 with an alpha of 0.05 and a power of 0.80 using the Tests for One Proportion procedure. (PASS 2021).

To evaluate the capabilities of ENDOANGEL-MM and the endoscopists, accuracy, sensitivity, and specificity for diagnosing gastric neoplasms were calculated for all the tests mentioned. The McNemar test was used to compare the accuracy, sensitivity, and specificity of Models. The Mann–Whitney *U* test was used to compare the accuracy of endoscopists and Models. $P < 0.05$ was considered statistically significant.

Role of the funding source

The funder had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

Image test by images and image-pairs

The diagnosis of Models 1 & 2 are based on images, while Models 3, 4 & 5 are on image pairs.

In dataset 2, the overall accuracy of Model 1, which performs diagnosis under WL, was 68.82% [95% CI 63.16–73.97%], with a sensitivity of 72.80% [95% CI 64.00–80.13%] and specificity of 66.06% [95% CI 58.54–72.85%]. The diagnosis of Model 2 is based on WM data, and its accuracy is significantly higher than that of Model 1 (78.44% [95% CI 74.51–81.91%] vs 68.81%, $P < 0.001$), but the advantage is mainly in specificity (90.78% [95% CI 86.83–93.63%] vs 66.06%, $P < 0.001$) rather than sensitivity (60.20% [95% CI 53.13–66.88%] vs 72.80%).

When integrated Model 1 and Model 2 at the task level, Model 3 mildly improved in diagnostic accuracy 79.03% [95% CI 76.79–81.11%], with a sensitivity of 83.24% [95% CI 80.26–85.85%] and specificity of 74.78% [95% CI 71.38–77.90%]. While the Model 4 achieved the best performance, slightly better than Model 3 in accuracy (82.11% [95% CI 79.99–84.05%] vs 79.03%), sensitivity (85.67% [95% CI 82.90–88.05%] vs 83.24%), and specificity (78.22% [95% CI 74.89–81.22%] vs 74.78%). The overall accuracy, sensitivity and specificity of Model 5 are 79.33% [95% CI 77.10–81.40%], 76.97% [95% CI 73.67–79.97%] and 81.71% [95% CI 78.62–84.44%].

Image test by lesions

As for lesions level, the results are consistent with those at the images and image-pairs. Models1 & 2 determine that a lesion is neoplastic when at least one image of the lesion is neoplastic. Similarly, Model 3 judges a lesion as neoplastic when at least one WM image and one WL image are simultaneously determined to be neoplastic. Models 4 & 5 identify neoplastic lesions by integrating the results of all image pairs.

The accuracy, sensitivity and specificity of Model 1 are 58.65% [95% CI 49.04–67.64%], 82.76% [95% CI 65.45–92.40%] and 49.33% [95% CI 38.33–60.40%], while those of Model 2 are better than Model 1 (78.85% [95% CI 70.05–85.60%], 86.21% [95% CI 69.44–94.50%] and 76.00% [95% CI 65.22–84.25%]) ($P = 0.0053$). The accuracy and specificity of Model 3 were slightly higher than that of Model 2 (82.69% [95% CI 74.29–88.76%] vs 78.85% [95% CI 70.05–85.60%]; 82.67% [95% CI 72.57–89.58%] vs 76.00% [95% CI 65.22–84.25%]), but the sensitivity is slightly lower (82.76% [95% CI 65.45–92.40%] vs 86.21% [95% CI 69.44–94.50%]). Model 5 outperformed Model 3 in accuracy (84.62% [95% CI 76.47–90.31%] vs 82.69% [95% CI 74.29–88.76%]) and specificity (85.33% [95% CI 75.61–91.61%] vs 82.67% [95% CI 72.57–89.58%]), while achieved the same sensitivity (82.76% [95% CI 65.45–92.40%]). Model 4 remains best performance with the accuracy of 86.54% [95% CI 78.67–91.81%], sensitivity of 89.66% [95% CI 73.62–96.42%], and specificity of 85.33% [95% CI 75.61–91.61%].

Given the above, the diagnostic ability at the lesion level is similar to those at the images & image-pairs level. Either at the images & image-pairs level or lesions level, integration at the feature layer (Model 4) achieved the best performance among models (Table 1).

Retrospective video test

For single-modal data, Model 2 performed significantly better than Model 1 in accuracy and specificity (85.00% [95% CI 76.72–90.69%] vs 45.00% [95% CI 35.61–54.76%], $P < 0.001$; 85.71% [95% CI 76.20–91.83%] vs 32.47% [95% CI 23.06–43.54%], $P < 0.001$), with a slightly lower sensitivity (82.61% [95% CI 62.86–93.02%] vs 86.96% [95% CI 67.88–95.46%]). For multi-modal data, Model 4 exceeds Model 3 and Model 5 in accuracy (90.00 vs 79.00%, 83.00%), sensitivity (95.65 vs 91.30%, 91.30%) and specificity (88.31 vs 75.32%, 80.52%), and exceeds the performance of single-modal models (vs Model 1: $P < 0.001$, vs Model 2: $P = 0.180$). These suggest that multi-modal models could achieve better diagnostic performance than single-modal models in both image and retrospective video tests.

Additionally, Model 1, Model 3, and Model 5 reached higher sensitivity in the retrospective video test than in the image test with the expense of some specificity reduction, while Model 4 improves both sensitivity and specificity at the same time. Integration at the feature level seems to provide greater robustness of the model.

Prospective video test

Performance of Models in prospective video test is similar to those in retrospective video test. Model 1 reached a satisfactory sensitivity of 93.75% [95% CI 71.67–98.89%], but the specificity and accuracy were unsatisfactory (23.91% [95% CI 13.91–37.93%], 41.94% [95% CI 30.48–54.34%]). The sensitivity, specificity and accuracy of Model 2 were 93.75% [95% CI 71.67–98.89%], 63.04% [95% CI 48.60–75.47%] and 70.97% [95% CI 58.71–80.78%]. For multi-modal data, Model 4 achieved the best performance as before, with a sensitivity of 93.75% [95% CI 71.67–98.89%], specificity of 93.48% [95% CI 82.50–97.76%] and accuracy of 93.55% [95% CI 84.55–97.46%], while those of Model 3 and Model 5 were 93.75% [95% CI 71.67–98.89%], 71.74% [95% CI 57.45–82.68%], 77.42% [95% CI 65.60–86.05%] and 93.75% [95% CI 71.67–98.89%], 91.30% [95% CI 76.96–95.27%], 91.94% [95% CI 82.48–96.51%] (Table 2).

Man–machine comparison

We conducted a man–machine contest among Model 2, Model 3, Model 4, Model 5, and 6 endoscopists on video test set. For WM single-modal data, the sensitivity and specificity of Model 2 in the video test set were 82.61%

Table 1 Performance comparison among Models on image testset (Dataset 2)

	Model 1	Model 2	Model 3	Model 4	Model 5
	Images		Image-Pairs		
Sensitivity (95% CI), %	72.80 (64.00–80.13)	60.20 (53.13–66.88)	83.24 (80.26–85.85)	85.67 (82.90–88.05)	76.97 (73.67–79.97)
Specificity (95% CI), %	66.06 (58.54–72.85)	90.78 (86.83–93.63)	74.78 (71.38–77.90)	78.22 (74.89–81.22)	81.71 (78.62–84.44)
Accuracy (95% CI), %	68.82 (63.16–73.97)	78.44 (74.51–81.91)	79.03 (76.79–81.11)	82.11 (79.99–84.05)	79.33 (77.10–81.40)
	Lesions				
Sensitivity (95% CI), %	82.76 (65.45–92.40)	86.21 (69.44–94.50)	82.76 (65.45–92.40)	89.66 (73.62–96.42)	82.76 (65.45–92.40)
Specificity (95% CI), %	49.33 (38.33–60.40)	76.00 (65.22–84.25)	82.67 (72.57–89.58)	85.33 (75.61–91.61)	85.33 (75.61–91.61)
Accuracy (95% CI), %	58.65 (49.04–67.64)	78.85 (70.05–85.60)	82.69 (74.29–88.76)	86.54 (78.67–91.81)	84.62 (76.47–90.31)

Model 1: White light (WL) images & lesions, Model 2: Weak magnification (WM) images & lesions, Model 3: task-level integration of WL and WM image-pairs & lesions, Model 4: feature-level integration of WL and WM image-pairs & lesions, Model 5: input-level integration of WL and WM image-pairs & lesions

Table 2 Performance comparison among Models on retrospective and prospective video testset (Dataset 3 and Dataset 4)

	Model 1	Model 2	Model 3	Model 4	Model 5
Dataset 3 (retrospective)					
Sensitivity (95% CI), %	86.96 (67.88–95.46)	82.61 (62.86–93.02)	91.30 (73.20–97.58)	95.65 (79.01–99.23)	91.30 (73.20–97.58)
Specificity (95% CI), %	32.47 (23.06–43.54)	85.71 (76.20–91.83)	75.32 (64.64–83.59)	88.31 (79.25–93.73)	80.52 (70.32–87.82)
Accuracy (95% CI), %	45.00 (35.61–54.76)	85.00 (76.72–90.69)	79.00 (70.02–85.83)	90.00 (82.56–94.48)	83.00 (74.45–89.11)
Dataset 4 (prospective)					
Sensitivity (95% CI), %	93.75 (71.67–98.89)	93.75 (71.67–98.89)	93.75 (71.67–98.89)	93.75 (71.67–98.89)	93.75 (71.67–98.89)
Specificity (95% CI), %	23.91 (13.91–37.93)	63.04 (48.60–75.47)	71.74 (57.45–82.68)	93.48 (82.50–97.76)	91.30 (76.96–95.27)
Accuracy (95% CI), %	41.94 (30.48–54.34)	70.97 (58.71–80.78)	77.42 (65.60–86.05)	93.55 (84.55–97.46)	91.94 (82.48–96.51)

Model 1: White light (WL) images & lesions, Model 2: Weak magnification (WM) images & lesions, Model 3: task-level integration of WL and WM image-pairs & lesions, Model 4: feature-level integration of WL and WM image-pairs & lesions, Model 5: input-level integration of WL and WM image-pairs & lesions

[95% CI 62.86–93.02%], 85.71% [95% CI 76.20–91.83%], better than the average of six endoscopists (66.67% [95% CI 58.45–73.99%], $P=0.107$; 73.17% [95% CI 68.94–77.00%], $P=0.007$).

For multi-modal data, Model 3 performed better than the average of endoscopists in accuracy (79.00% [95% CI 70.02–85.83%] vs 76.17% [95% CI 72.60–79.41%]) and sensitivity (91.30% [95% CI 73.20–97.58%] vs 75.36% [95% CI 67.55–81.80%]), with a slightly lower specificity (75.32% [95% CI 64.64–83.59%] vs 76.41% [95% CI 72.33–80.05%]). Model 4 outperformed the average of endoscopist significantly in accuracy, sensitivity and specificity (90.00% [95% CI 82.56–94.48%] vs 76.17% [95% CI 72.60–79.41%], $P = 0.002$; 95.65% [95% CI 79.01–99.23%] vs 75.36% [95% CI 67.55–81.80%], $P = 0.002$; 88.31% [95% CI 79.25–93.73%] vs 76.41% [95% CI 72.33–80.05%], $P = 0.040$), exceeding the performance of expert endoscopists (87.00% [95% CI 81.63–90.97%], 80.44% [95% CI 66.82–89.35%], 88.97% [95% CI 83.03–92.99%]). Model 5 made fusion at the input layer and achieved the sensitivity and specificity of 91.30% [95% CI 73.20–97.58%] and 80.52% [95% CI 70.32–87.82%].

Either endoscopists or machine performs better when multi-modal data were available. AI achieves better results than endoscopists in both single-modal and multi-modal modes (Table 3).

Table 3 Diagnostic ability of models compared with endoscopists' performance with and without AI's assistance

	WM single-modal data			WM &WL multi-modal data			Endoscopists' performance with AI's assistence		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
Model 2	85.00	82.61	85.71	–	–	–	–	–	–
Model 4	–	–	–	90.00	95.65	88.31	–	–	–
Expert									
Expert 1	84.00	86.96	83.17	89.00	73.91	93.51	91.00	86.95	92.20
Expert 2	81.00	69.57	84.41	85.00	86.96	84.42	87.00	95.65	84.42
Average of expert	82.50	78.27	83.79	87.00	80.44	88.97	89.00	91.30	88.31
Non-expert									
Senior 1	79.00	56.52	85.71	81.00	69.57	84.42	88.00	69.57	93.51
Senior 2	65.00	82.61	59.74	75.00	78.26	74.03	86.00	95.65	83.17
Junior 1	62.00	52.17	64.94	69.00	60.86	71.42	86.00	82.61	87.01
Junior 2	59.00	52.17	61.03	58.00	82.60	50.64	83.00	91.30	80.52
Average of non-expert	66.25*	60.80*	67.86*	70.75^	72.82^	70.13^	85.75$^{\#}$	84.78	86.05
Average of total	71.67**	66.67	73.17**	76.17^^	75.36^^	76.41^	86.83	86.96	86.81

WM weak magnification, *WL* white light

*Compared with Model 2, ^compared with Model 4, $^{\#}$compared with endoscopists' performance without AI's assistance

*/^/$^{\#}$ Significant at 5% level. **/^^/$^{\#\#}$ Significant at 1% level. ***/^^^/$^{\#\#\#}$ Significant at 0.1% level

The performance of endoscopists with and without ENDOANGEL-MM's assistance

Since Model 4 has better results than Model 3 and Model 5 in both image tests and video tests, the results of Model 4 are provided to endoscopists for assistance diagnosis. With the assistance of the ENDOANGEL-MM (Model 4), the endoscopists' accuracy improved (86.83% [95% CI 78.82–91.22%] vs 76.17% [95% CI 61.64–79.13%], $P = 0.054$). Junior endoscopists perform comparable to experts with the support of the ENDOANGEL-MM (84.50% [95% CI 76.15–90.30%] vs 89.00% [95% CI 81.37–93.75%], $P = 0.121$). With the aid of the ENDOANGEL-MM, the time taken by endoscopists to make a diagnosis can be effectively reduced (30.67 vs 65.17 min, $P < 0.05$) (Table 3).

Subgroup analysis

When data of the advanced gastric cancer were excluded, Model 4 remains the best on image test, retrospetive and prospective video test. In the man–machine comparison, Model 2 outperformed endoscopists on WM single-modal data (84.04% [95% CI 75.32–90.08%] vs 69.86% [95% CI 65.95–73.50%], $P = 0.002$) and Model 4 outperformed endoscopists on multi-modal data (89.36% [95% CI 81.51–94.12%] vs 74.64% [95% CI 70.90–78.07%], $P = 0.002$) significantly. With the assistance of the ENDOANGEL-MM (Model 4), the endoscopists' accuracy improved (86.83% [95% CI 82.88–88.61%] vs 74.64% [95% CI 70.90–78.07%], $P = 0.054$). Junior endoscopists perform comparable to experts with the support of the ENDOANGEL-MM (83.51% [95% CI 77.55–88.13%] vs 88.30% [95% CI 82.92–92.15%], $P = 0.121$). (Tables S2, S3, S4).

Discussion

Early diagnosis of gastric neoplasms under endoscope is crucial but remains challenging. In this study, an artificial intelligence model was developed for identifying gastric neoplasms under WL and WM dual-modal modes and was validated in image test set, retrospective video test, prospective patients, man–machine comparison, and assisted diagnosing test. The system achieved well diagnostic efficacy and performed similar to experts ($P = 0.102$).

WL and ME have long been used for the diagnosis of gastric neoplasms (high-risk gastric lesions), and the advent of the 'near-focus (NF)' mode of Olympus provides a good opportunity for weak magnification mode to become a stand-alone diagnostic module. Under the NF mode, the endoscopists are able to observe capillaries and tissue structures, and can obtain high-quality images. The NF mode introduces visibility in the routine inspection, therefore it is widely applied in mucosal lesions diagnosing [13]. Not only NF mode, but also the blue laser imaging (BLI) mode of Fuji contains a certain magnification interval. To increase the applicability, we incorporated the image data of magnifying narrow band imaging (M-NBI) and magnifying blue laser imaging (M-BLI) with the same magnification range as NF mode to construct the weak magnification dataset.

Previous studies have been paying much attention to the diagnosis of early gastric cancer and gastric neoplasms using artificial intelligence. Yoon et al. developed a system to optimize early gastric cancer detection and depth prediction, which achieved a sensitivity of 91.0% and a specificity of 97.60% [19]. Ueyama et al. constructed a system to diagnose early gastric cancer under magnifying endoscopy and achieved an accuracy of 98.7% [20]. However, previous studies have mainly used WL or ME single-mode data, and the use of WM mode has not been explored. Diagnoses based on single-modal data may result in the omission of focal features. The guidelines also recommended the use of multi-modal light source combined with high definition WL endoscopy to diagnose gastric neoplasms [18]. One recent research confirmed that AI combining WL and IEE together achieved better performance in diagnosing invasion depth of colorectal cancer, compared with WL and IEE solely [21]. There were few studies reporting multi-modal-based AI systems, and the diagnostic ability of different multi-modal models has not been explored yet either. In this study, we first developed various models based on single-modal or multi-modal data, validated the diagnosing ability of the Models at different levels, and compared the performance of different multi-modal models.

When considering whether the diagnostic results of AI could replace biopsy, we hold the view that AI would not replace biopsy but could assist in it and potentially reduce unnecessary biopsy. A guideline revealed that endoscopic characterization with IEE avoids unnecessary biopsies for upper GI superficial lesions [22]. In this study, the AI systems outperformed human in diagnosing early gastric cancer, which indicated that the unnecessary biopsies could be potentially reduced with the assistance of AI.

The other strength of our system is the ability to process videos and give predictions in real time, which was essential for an AI system. Endoscopic examination is dynamic and real-time in clinical practice. Receiving feedback from AI during real-time operation could improve interaction between endoscopists and machine. High-quality human–machine interaction may prevent endoscopists' individual pitfalls and improve patients outcome [23]. In assisted diagnosing test, the performance of junior endoscopists was inferior to that of experts without AI assistance (82.50 vs 60.50%), while comparable to that of experts with AI assistance (84.50 vs 89.00%). Notably, the performance of junior endoscopists and experts improved with AI assistance.

These findings implied that ENDOANGEL-MM could help effectively in the detection of gastric neoplasms.

When assisted with an AI system, the endoscopists' decision on their final diagnosis may be influenced by multiple factors. Relevant study revealed that age ($P = 0.013$), professional title ($P = 0.001$), and the duration of using AI ($P = 0.000$) influence endoscopists' acceptance of AI significatly, and may affect the decision of endoscopists [24]. In this study, before the assisted diagnosing test, we informed endoscopists that they could make diagnosis based on their own experience and the results of AI. The results of assisted diagnosing test showed that even if the final performances of juniors and experts are comparable, all of the wrong diagnoses made by the juniors were also made by the AI (20/20), while some of the wrong diagnoses of experts were made by themselves (12/20). This implies that, when faced with contradictions in diagnostic judgment with AI, junior endoscopists appear to be more receptive to AI's results, while experts are more likely to trust on their own, which is consistant with previous study. Study proved that AI can assist endoscopists detect more positive lesions meanwhile preventing unnecessary biopsies [25]. Therefore, for AI systems that have already proven good effect in clinical trials, the endoscopists may take the AI recommendations into consideration. Sometimes AI may make mistakes due to noises. However, obvious false positive and false negative predictions of AI can be easily identified and be compensated for by an autonomous diagnosis by the endoscopist.

Generally, the advanced gastric cancer is easier to detect because of its typical characteristics, and it is worth dicussing whether advanced cancers should be included in the develepment of AI systems. We hold the view that if advanced gastric cancer was excluded, the AI model may not judge advanced gastric cancer as high-risk in clinical practice, which may limit the applicability of AI system and reduce human trust. A study also stated that when only cases of high-grade dysplasia or early cancer are included, the system may miss an advanced cancer [23]. Therefore, in consideration to avoid spectrum bias, we included the advanced cancer in our study. Notably, we made sub-analysis of the testing process by excluding the advanced gastric cancer from the test sets to fully evaluate the models' performance and superiority on early gastric neoplasms. The results showed that AI was superior to human in diagnosing gastric neoplasms regardless of whether advanced gastric cancer was included, and the model with multi-modal fusion on feature level was still the best model among AI models, which were consistent with the conclusions of the main analysis.

There are several limitations to our approach. First, only single-center data from the Renmin Hospital of Wuhan University were used in this study, which may increase the risk of bias, so we will enhance and test the system using multicenter data in the future. Second, although the performance of the ENDOANGEL-MM was fully tested in images, videos and prospective patients, clinical trials evaluating its clinical effect should be further conducted.

In conclusion, this study is the first to investigate the ability of a multi-modal AI system to diagnose gastric neoplasms in WL combined with WM mode. The ENDOANGEL-MM could effectively assist endoscopists in detecting gastric neoplasms and provide new ideas for the construction of subsequent artificial intelligence systems.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10120-022-01358-x.

Acknowledgements This study was financially supported by Health Commission of Hubei Province (Innovation Team Project of Health Commission of Hubei Province, grant no. WJ202C003, to Honggang Yu) and Wuhan University (The Fundamental Research Funds for the Central Universities, grant no. 2042021kf0084, to Lianlian Wu).

Author contributions HGY and LLW conceived and designed the study; YXL, CJL, YCD, DC, and WXD collected and reviewed images; HLD, ZHD, CJL, YJZ collected, collated and analyzed the data; HLD, XT, JXW, XQZ, CXZ collected and edited the video clips; HLD, ZHD wrote the manuscript; LLW and ZHD revised the manuscript; HGY performed extensive editing of the manuscript; all authors reviewed and approved the final manuscript for submission. All authors were involved in data acquisition, general design of the trial, interpretation of the data, and critical revision of the manuscript. We ensured that all the authors had access to all the raw data sets.

Data availability Individual deidentified participant data that underline the results reported in this article will be shared after article publication. To gain access, data requesters could contact the corresponding author.

Declarations

Conflict of interest There is no conflict of interest.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemalt A, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 Countries. CA Cancer J Clin. 2021;71(3):209–49.
2. Smyth EC, Nilsson M, Grabsch HI, Grieken NC, Lordic F. Gastric cancer. Lancet. 2020;396(10251):635–48.
3. Niu P, Zhao L, Wu H, Zhao D, Chen Y. Artificial intelligence in gastric cancer: application and future perspectives. World J Gastroenterol. 2020;26(36):5408–19.
4. Fitzmaurice C, Abate D, Abbasi N, Abbastabar H, Foad A, Omar A, et al. Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 29 cancer groups, 1990 to 2017: a systematic analysis for the Global Burden of Disease Study. JAMA Oncol. 2019;5(12):1749–68.
5. Banks M, Graham D, Jansen M, Gotoda T, Coda S, Pietro M, et al. British Society of Gastroenterology guidelines on the diagnosis and management of patients at risk of gastric adenocarcinoma. Gut. 2019;68(9):1545–75.

6. Pimentel-Nunes P, Libânio D, Marcos-Pinto R, Areia M, Leja M, Esposito G, et al. Management of epithelial precancerous conditions and lesions in the stomach (MAPS II): European Society of Gastrointestinal Endoscopy (ESGE), European Helicobacter and Microbiota Study Group (EHMSG), European Society of Pathology (ESP), and Sociedade Portuguesa de Endoscopia Digestiva (SPED) guideline update 2019. Endoscopy. 2019;51(4):365–88.
7. Wang F, Zhang X, Li Y, Tang L, Qu X, Ying J, et al. The Chinese Society of Clinical Oncology (CSCO): clinical guidelines for the diagnosis and treatment of gastric cancer, 2021. Cancer Commun (Lond). 2021;41(8):747–95.
8. Zhang X, Li M, Chen S, Hu J, Guo Q, Liu R, et al. Endoscopic screening in Asian countries is associated with reduced gastric cancer mortality: a meta-analysis and systematic review. Gastroenterology. 2018;155(2):347–54.
9. Zhang Q, Wang F, Chen Z, Wang Z, Zhi F, Liu S, et al. Comparison of the diagnostic efficacy of white light endoscopy and magnifying endoscopy with narrow band imaging for early gastric cancer: a meta-analysis. Gastric Cancer. 2016;19(2):543–52.
10. Undo N, Yao K. Endoluminal diagnosis of early gastric cancer and its precursors: bridging the gap between endoscopy and pathology. Adv Exp Med Biol. 2016;908:293–316.
11. Song MJ, Ang TL. Early detection of early gastric cancer using image-enhanced endoscopy: current trends. Gastrointestinal Intervention. 2014;3(1):1–7.
12. He X, Wu L, Dong Z, Gong D, Jiang X, Zhang H, et al. Real-time use of artificial intelligence for diagnosing early gastric cancer by magnifying image-enhanced endoscopy: a multicenter, diagnostic study. Gastrointest Endosc. 2022;95(4):671–8.
13. Kakushima N, Yoshida N, Doyama H, Yano T, Horimatsu T, Uedo N, et al. Near-focus magnification and second-generation narrow-band imaging for early gastric cancer in a randomized trial. J Gastroenterol. 2020;55(12):1127–37.
14. An P, Yang D, Wang J, Wu L, Zhou J, Zeng Z, et al. A deep learning method for delineating early gastric cancer resection margin under chromoendoscopy and white light endoscopy. Gastric Cancer. 2020;23(5):884–92.
15. Ling T, Wu L, Fu Y, Xu Q, An P, Zhang J, et al. A deep learning-based system for identifying differentiation status and delineating the margins of early gastric cancer in magnifying narrow-band imaging endoscopy. Endoscopy. 2021;53(5):469–77.
16. Wu L, He X, Liu M, Xie H, An P, Zhang J, et al. Evaluation of the effects of an artificial intelligence system on endoscopy quality and preliminary testing of its performance in detecting early gastric cancer: a randomized controlled trial. Endoscopy. 2021;53(12):1199–207.
17. He X, Deng Y, Fang L, Peng Q. Multi-modal retinal image classification with modality-specific attention network. IEEE Trans Med Imaging. 2021;40(6):1591–602.
18. Pedro PN, Diogo L, Ricardo MP, Areia M, Leja M, Esposito G, et al. Management of epithelial precancerous conditions and lesions in the stomach (MAPS II): European Society of Gastrointestinal Endoscopy (ESGE), European Helicobacter and Microbiota Study Group (EHMSG), European Society of Pathology (ESP), and Sociedade Portuguesa de Endoscopia Digestiva (SPED) guideline update 2019. Endoscopy. 2019;51(4):365–88.
19. Yoon H, Kim S, Kim J, Keum J, Oh S, Jo J, et al. A lesion-based convolutional neural network improves endoscopic detection and depth prediction of early gastric cancer. J Clin Med. 2019;8(9):1310.
20. Ueyama H, Kato Y, Akazawa Y, Yatagai N, Komoti H, Takeda T, et al. Application of artificial intelligence using a convolutional neural network for diagnosis of early gastric cancer based on magnifying endoscopy with narrow-band imaging. J Gastroenterol Hepatol. 2021;36(2):482–9.
21. Lu Z, Xu Y, Yao L, Zhou W, Gong W, Yang G, et al. Real-time automated diagnosis of colorectal cancer invasion depth using a deep learning model with multimodal data (with video). Gastrointest Endosc. 2022;95(6):1186-94.e3.
22. Chiu P, Uedo N, Singh R, Gotoda T, Ng E, Yao K, et al. An Asian consensus on standards of diagnostic upper endoscopy for neoplasia. Gut. 2019;68(2):186–97.
23. Sharma P, Hassan C. Artificial intelligence and deep learning for upper gastrointestinal neoplasia. Gastroenterology. 2022;162(4):1056–66.
24. Tian L, Zhang Z, Long Y, Tang A, Deng M, Long X, et al. Endoscopists' acceptance on the implementation of artificial intelligence in gastrointestinal endoscopy: development and case analysis of a scale. Front Med (Lausanne). 2022;9: 760634.
25. Wu L, Shang R, Sharma P, Zhou W, Liu J, Yao L, et al. Effect of a deep learning-based system on the miss rate of gastric neoplasms during upper gastrointestinal endoscopy: a single-centre, tandem, randomised controlled trial. Lancet Gastroenterol Hepatol. 2021;6(9):700–8.