Development and validation of a feature extraction-based logical anthropomorphic diagnostic system for early gastric cancer: A case-control study

Jia Li,[a,b,c,1] Yijie Zhu,[a,b,c,1] Zehua Dong,[a,b,c,1] Xinqi He,[a,b,c] Ming Xu,[a,b,c] Jun Liu,[a,c,d] Mengjiao Zhang,[a,b,c] Xiao Tao,[a,b,c] Hongliu Du,[a,b,c] Di Chen,[a,b,c] Li Huang,[a,b,c] Renduo Shang,[a,b,c] Lihui Zhang,[a,b,c] Renquan Luo,[a,b,c] Wei Zhou,[a,b,c] Yunchao Deng,[a,b,c] Xu Huang,[a,b,c] Yanxia Li,[a,b,c] Boru Chen,[a,b,c] Rongrong Gong,[a,b,c] Chenxia Zhang,[a,b,c] Xun Li,[a,b,c] Lianlian Wu,[a,b,c]* and Honggang Yu [a,b,c]*

[a]Department of Gastroenterology, Renmin Hospital of Wuhan University, 99 Zhangzhidong Road, Wuhan, Hubei 430060, PR China
[b]Hubei Key Laboratory of Digestive System, Renmin Hospital of Wuhan University, Wuhan, Hubei 430060, PR China
[c]Hubei Provincial Clinical Research Center for Digestive Disease Minimally Invasive Incision, Renmin Hospital of Wuhan University, Wuhan, Hubei 430060, PR China
[d]Nursing Department of Renmin Hospital of Wuhan University, Wuhan, Hubei, 430060, PR China

eClinicalMedicine
2022;46: 101366
Published online 30 March 2022
https://doi.org/10.1016/j.eclinm.2022.101366

Summary

Background Prompt diagnosis of early gastric cancer (EGC) is crucial for improving patient survival. However, most previous computer-aided-diagnosis (CAD) systems did not concretize or explain diagnostic theories. We aimed to develop a logical anthropomorphic artificial intelligence (AI) diagnostic system named ENDOANGEL-LA (logical anthropomorphic) for EGCs under magnifying image enhanced endoscopy (M-IEE).

Methods We retrospectively collected data for 692 patients and 1897 images from Renmin Hospital of Wuhan University, Wuhan, China between Nov 15, 2016 and May 7, 2019. The images were randomly assigned to the training set and test set by patient with a ratio of about 4:1. ENDOANGEL-LA was developed based on feature extraction combining quantitative analysis, deep learning (DL), and machine learning (ML). 11 diagnostic feature indexes were integrated into seven ML models, and an optimal model was selected. The performance of ENDOANGEL-LA was evaluated and compared with endoscopists and sole DL models. The satisfaction of endoscopists on ENDOANGEL-LA and sole DL model was also compared.

Findings Random forest showed the best performance, and demarcation line and microstructures density were the most important feature indexes. The accuracy of ENDOANGEL-LA in images (88.76%) was significantly higher than that of sole DL model (82.77%, $p = 0.034$) and the novices (71.63%, $p<0.001$), and comparable to that of the experts (88.95%). The accuracy of ENDOANGEL-LA in videos (87.00%) was significantly higher than that of the sole DL model (68.00%, $p<0.001$), and comparable to that of the endoscopists (89.00%). The accuracy (87.45%, $p<0.001$) of novices with the assistance of ENDOANGEL-LA was significantly improved. The satisfaction of endoscopists on ENDOANGEL-LA was significantly higher than that of sole DL model.

Interpretation We established a logical anthropomorphic system (ENDOANGEL-LA) that can diagnose EGC under M-IEE with diagnostic theory concretization, high accuracy, and good explainability. It has the potential to increase interactivity between endoscopists and CADs, and improve trust and acceptability of endoscopists for CADs.

Funding This work was partly supported by a grant from the Hubei Province Major Science and Technology Innovation Project (2018-916-000-008) and the Fundamental Research Funds for the Central Universities (2042021kf0084).

Keywords: Feature extraction; Logical anthropomorphic artificial intelligence; Early gastric cancer; Magnifying image enhanced endoscopy

*Correspondeing authors at: Department of Gastroenterology, Renmin Hospital of Wuhan University, 99 Zhangzhidong Road, Wuhan, Hubei 430060, PR China.
E-mail addresses: wu_leanne@163.com (L. Wu), yuhonggang@whu.edu.cn (H. Yu).

[1] These authors contributed equally to this work.

Research in context

Evidence before this study

We searched PubMed for papers published from Jan 1, 1999 to Dec 31, 2021, with the keywords "artificial intelligence" OR "deep learning" OR "machine learning" AND "early gastric cancer" AND "endoscopy". In recent years, great efforts have been made by researchers with the help of artificial intelligence (AI) in assisting early gastric cancer (EGC) diagnosis under magnifying image enhanced endoscopy (M-IEE). However, most AI models only output an answer "cancer" or "non-cancer" without details about how the diagnosis process was made, and cannot show the abstract diagnostic theories to endoscopists in a concrete and intuitive way.

Added value of this study

In this study, we developed and validated a logical anthropomorphic (LA) AI diagnostic system named ENDOANGEL-LA for EGCs under M-IEE, which is based on feature extraction combining quantitative analysis, deep learning (DL), and machine learning (ML). The performance of ENDOANGEL-LA with diagnostic theories concretization and good explainability was better than sole DL models and comparable to that of experts. To our knowledge, this is the first study developing a logical anthropomorphic AI system based on feature extracting to diagnose EGC with both good accuracy and explainability, and to concretize abstract diagnostic theories.

Implications of all the available evidence

AI systems such as ENDOANGEL-LA have great potential in assisting the diagnosis of EGC. Our system was able to diagnose EGC under M-IEE with diagnostic theories concretization, high accuracy, and good explainability. The system has the potential to increase interactivity between endoscopists and AI systems based on the features extracted, and improve the trust and acceptability of endoscopists on AI systems.

Introduction

Gastric cancer (GC) is the third leading cause of cancer-related death globally, with estimates of more than 750,000 deaths worldwide during 2020.[1–4] Patients diagnosed with early gastric cancer (EGC) have a 5-year survival rate of more than 90%, which decreases to less than 25% when evolving into the advanced stage.[5] Early detection of EGC by endoscopy is the prerequisite for timely endoscopic treatment and patients' welfare.[6–8] However, as EGCs usually show subtle changes of the mucosa, accurate diagnosis of EGC under white light endoscopy is difficult. As a result, the miss diagnosis rate of EGC is as high as 20–40%.[9]

Therefore, the magnifying image enhanced endoscopy (M-IEE), which can clearly show the microstructures (MS) and microvessels (MV) of the gastric mucosa, has been developed and widely applied to improve EGC diagnosis.[10] In real clinics, however, the performance among endoscopists varies greatly because the diagnosis under M-IEE requires extensive experience and thorough knowledge.[11,12] This is because the current diagnostic theories of EGC under M-IEE are mainly an abstract generalization of MS and MV, which has intuitiveness, uncertainty, and fuzziness. A study involving 395 endoscopists in 77 medical institutions showed that based on the diagnostic theories, the accuracy of different endoscopists in diagnosing EGC under M-IEE fluctuated between 40% and 85%,[13] which seriously affected the detection of early gastric cancer.

To assist in EGC diagnosis under M-IEE in real time, great efforts have been made by researchers with the help of artificial intelligence (AI). Hu, et al. collected 1777 M-IEE images for constructing a computer-aided-diagnosis (CAD) model and earned a sensitivity of 0.792.[12] Yusuke Horiuchi et al. established a CAD system using 2570 images with an accuracy of 85.1% in videos.[14] Although previous studies have confirmed that AI has great potential in assisting the diagnosis of EGC, most AI models only output an answer "cancer" or "non-cancer" without details about how the diagnosis process was made,[15] and cannot show the abstract diagnostic theories to endoscopists in a concrete and intuitive way. As a result, it is difficult for endoscopists to learn from the models and find out the cause of the errors and ways to avoid them,[15] which will greatly limit the clinical applications of AI systems.

In the present study, we developed a feature extraction-based logical anthropomorphic diagnostic system named ENDOANGEL-LA (logical anthropomorphic) for EGC under M-IEE combined with prior knowledge, quantitative analysis, deep learning (DL), and machine learning (ML). The performance of ENDOANGEL-LA for diagnosing EGC under M-IEE was tested using still images, prospective videos and further compared with endoscopists of different levels. A comparison between the ENDOANGEL-LA and the sole DL models of different training sets on their performance and the satisfaction of endoscopists was also conducted. To the best of our knowledge, this is the first study to concretize abstract diagnostic theories with feature extraction and achieve diagnostic logic transparency of AI systems in the field of EGC diagnosis under M-IEE.

Methods

Datasets

This case-control study was done in Renmin Hospital of Wuhan University (RHWU), Wuhan, China. According to pathological results, the images were divided into the EGC group (case group) and the non-cancer group (control group). We estimated the accuracy of

ENDOANGEL-LA was 89% when diagnosing EGC in M-IEE images and the accuracy of sole DL model was 80% based on a previous pilot study. The estimated sample size of image test set was 250 with a type I error rate of 0.05 and power of 0.80. Therefore, we retrospectively collected 4667 images (EGC, 1950; noncancerous, 2717) from 1811 patients (EGC, 1042; noncancerous lesions, 769) between Nov 15, 2016 and May 7, 2019, which were derived from a database of our previous study.[16] To develop ENDOANGEL-LA, the images viewed at full magnification under M-IEE were included, and the images where the MS and MV were difficult to observe due to blurry, reflection, out of focus, etc. were excluded. Finally, 1897 images (EGC, 679; noncancerous, 1218) from 692 patients (EGC, 363; noncancerous lesions, 329) were enrolled. Enrolled images were randomly assigned to a training set (1630 images from 567 patients) and test set (267 images from 125 patients) by patient with a ratio of about 4:1. The workflow of this study is illustrated in Figure S1. The patient and lesion characteristics in test set are shown in Table S2.

The M-IEE was performed by a standard magnifying endoscopy [(EG-L590ZW; Fujifilm, Tokyo, Japan), (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)]. The resolution of all the images is 512×512 pixels.

Establishment of feature index base for EGC endoscopic diagnosis with prior knowledge

To provide ENDOANGEL-LA with sufficient prior knowledge to determine the feature indexes related to EGC diagnosis under M-IEE, the eligible studies published from January 1, 1999 to December 31, 2021 were searched by the keywords "Early gastric cancer", "Magnifying endoscopy" and "Diagnosis" in PubMed databases. A total of 203 pieces of published literatures were searched and assessed. Among them, one piece of duplicate literature was removed, and 38 pieces of literatures were excluded after screening of title and abstract because they are unrelated to EGC or magnifying endoscopy (*n* = 21) and are case reports (*n* = 17). Then, 124 pieces of literatures were excluded because the full texts were unavailable (*n* = 3) and unrelated to EGC diagnostic features (*n* = 121). Finally, 40 pieces of literatures were obtained. Based on the selected literature, two expert endoscopists and two algorithmic engineers jointly determined feature indexes related to EGC diagnosis. Then according to the performance of each feature index and the similarity between feature indexes, eleven feature indexes were finally determined for inclusion. (Figure S2)

Development of the ENDOANGEL-LA

Eleven feature indexes including seven quantitative feature indexes and four deep learning feature indexes were used to develop the ENDOANGEL-LA. The quantitative feature indexes included: (1) density of MS, (2) eccentricity of MS equivalent centroid, (3) diameter ratio of MV, (4) tortuosity of MV, (5) cyclization of MV, (6) spectral principal component information of gastric mucosal background color, (7) image entropy of S-channel in Hue-Saturation-Intensity (HSI) color space. The deep learning feature indexes included: (1) the arrangement of the M-IEE images, (2) the demarcation line of lesions, (3) the distribution of MV in the MV segmentation images, (4) the morphology of the lesions (including elevated, flat, and depressed). All the feature indexes are shown in Figure 1. The performance of all the feature indexes is described in Table S3 and Figures S3−6.

Five quantitative feature indexes were analyzed based on clear areas of the images, which is to eliminate the influence of bleeding, blurring, bubbles, etc. on the accuracy of MS and MV analysis. Details of the segmentation of clear areas, MS, and MV are described in the Supplementary Material and Figure 2. The definitions of each feature index are as follows (The diagnosis theories of EGC and the definition of corresponding quantitative indexes are described in detail in Table S4 and Figure 1):

(1) Density of MS: The ratio of the number of MS pixels in the MS segmentation image to the total number of pixels in the clear area. It was used to describe the distribution of MS.[17] The lower density of MS, the greater possibility of EGC.

(2) Eccentricity of MS equivalent centroid: The connected component analysis was used to extract the centroid of each MS in the MS segmentation image and the centroid of the clear area. The equivalent centroid of the MS segmentation image was obtained based on the centroid of each MS. The eccentricity of the MS equivalent centroid is the offset distance between the equivalent centroid of the MS segmentation image and the centroid of the clear area. It was used to describe the distribution of MS.[17] The further eccentricity of MS equivalent centroid, the greater possibility of EGC.

(3) Diameter ratio of MV: The connected component analysis was used to extract a single MV in the MV segmentation image, and the centerline of the MV was extracted based on the single MV. Calculate the diameter of every point on the MV. The representative value of diameter ratio of MV is the simple geometric mean of the ratio of the maximum diameter to the minimum diameter of all MV. It was used to describe the morphology of MV.[17] The further eccentricity of MS equivalent centroid, the greater possibility of EGC.

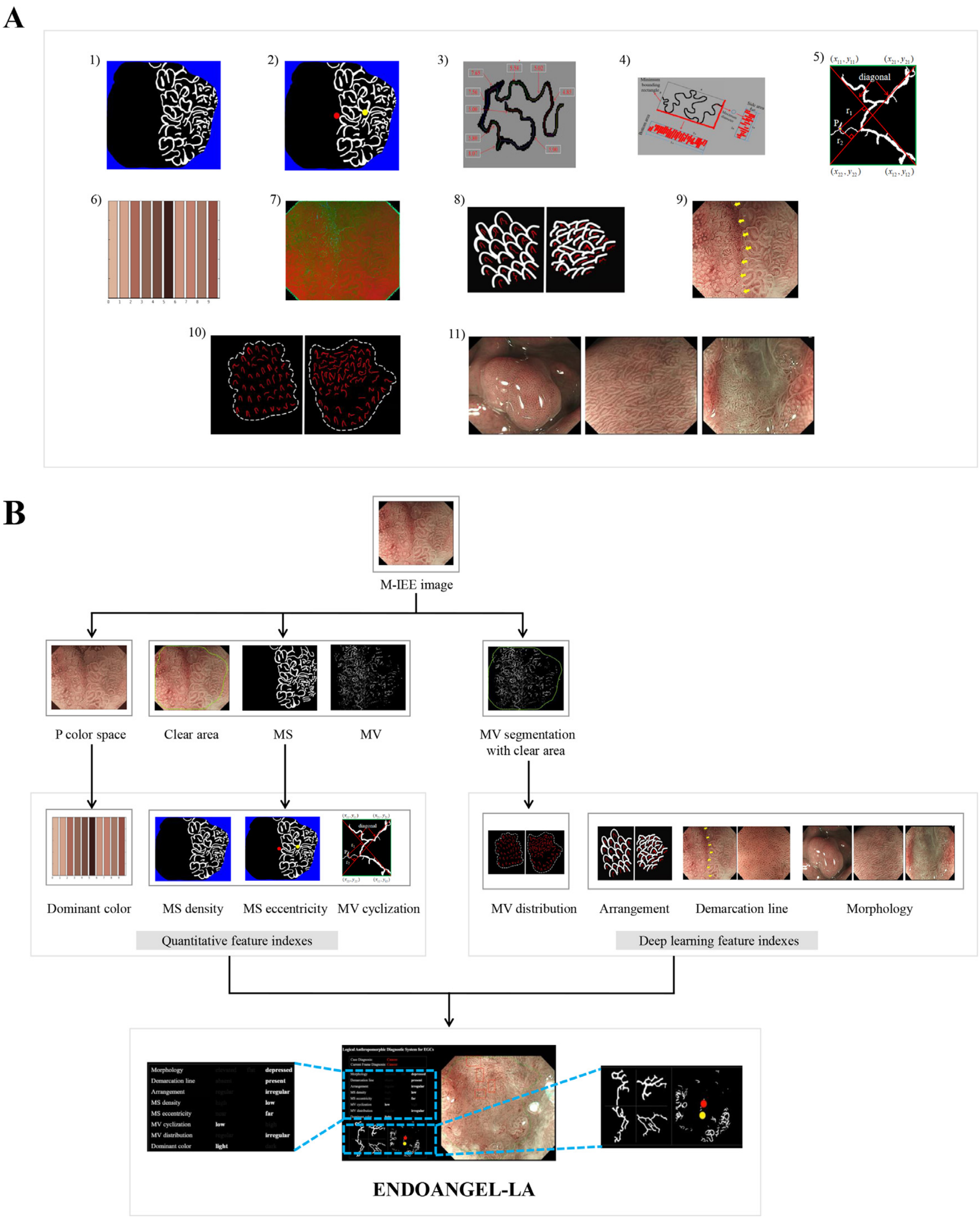

Figure 1. The schematic diagram of all feature indexes and the framework of developing ENDOANGEL-LA.

(A) Eleven feature indexes. (1) density of MS. (2) eccentricity of MS equivalent centroid: yellow dots represent the equivalent centroid of MS segmentation image, and red dots represent the centroid of the clear area. (3) diameter ratio of MV. (4) tortuosity of MV. (5) cyclization of MV. (6) ten main color features. (7) HSI color space. (8) the arrangement of the M-IEE images: the left image shows regular arrangement and the right image shows irregular arrangement. (9) the demarcation line of lesions: there is a demarcation line in the image (yellow arrows). (10) the distribution of MV in the MV segmentation images: the left image shows regular distribution and the right image shows irregular distribution. (11) the morphology of the lesions: the left image shows an elevated lesion, the middle image shows a flat lesion, and the right image shows a depressed lesion. (B) The framework of developing ENDOANGEL-LA. MS: microsurfaces, MV: microvessels, HSI: Hue-Saturation-Intensity, EGC: early gastric cancer, M-IEE: magnifying image-enhanced endoscopy, LA: logical anthropomorphic.

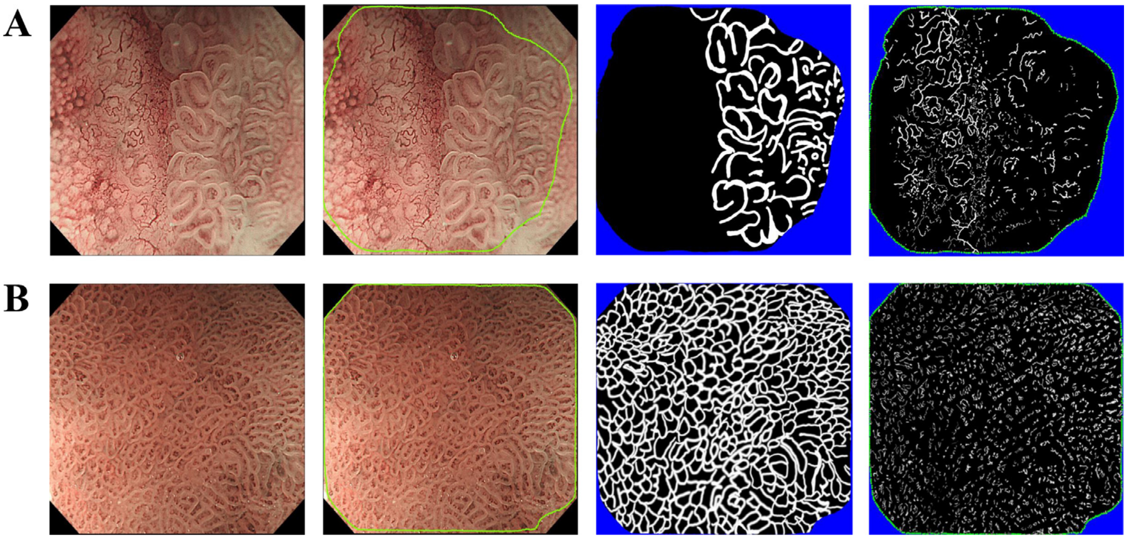

Figure 2. The segmentation of clear areas, MS and MV. (A) EGC images. (B) non-cancerous images.

The images from left to right are M-IEE image, M-IEE image with clear area segmentation, MS segmentation image, and the MV segmentation image. MS: microsurfaces, MV: microvessels, EGC: early gastric cancer, M-IEE: magnifying image-enhanced endoscopy.

(4) Tortuosity of MV: The minimum bounding cube was constructed with the minimum bounding rectangle and the maximum diameter of MV. The surface density was calculated according to the total number of pixels on the MV and the bottom area or side area of the minimum bounding cube, and the tortuosity coefficient of MV was obtained by weighting the surface density. The representative value of tortuosity of MV is the median of tortuosity coefficients of all MV. It was used to describe the morphology of MV.[17] The larger tortuosity coefficient of MV, the greater probability of EGC.

(5) Cyclization of MV: The two diagonals' area moment of inertia was calculated based on the minimum bounding rectangle and all the pixels on the MV. The cyclization coefficient of MV was obtained by weighting the area moment of inertia. The representative value of cyclization of MV is the mean of the ratio of the cyclization coefficients of all MV. It was used to describe the morphology of MV.[17] The smaller cyclization coefficient of MV, the higher cyclization of MV, and the greater probability of EGC.

(6) Spectral principal component information of gastric mucosal background color: Transform the image from Red-Green-Blue (RGB) color space to P color space, and extract ten main color features of the images in P color space. Then the average pixels of each color feature in the three channels were calculated, and the median of all average pixels is the representative value of spectral principal component information. It was used to describe the color of the gastric mucosa.[17] The smaller the representative value, the greater probability of EGC.

(7) Image entropy of S-channel in HSI color space: Transform the image from RGB color space to HSI color space, and calculate the image entropy in the S-channel.[17,18] It was used to describe the color of the gastric mucosa. The larger the image entropy, the greater probability of EGC.

Four models of the deep learning feature indexes were constructed with ResNet-50 using 1630 images and tested using 267 images. And the images were evaluated by 2 expert endoscopists for the arrangement of the M-IEE images, the demarcation line of lesions, the distribution of MV in the MV segmentation images, and the morphology of the lesions. If there was disagreement between the two endoscopists, a reassessment was carried out to reach a consensus.

All feature indexes were arranged and combined and input into the machine learning (ML) models, including random forest (RF), Gaussian Naive Bayes (GNB), k-Nearest Neighbor (KNN), logistic regression (LR), decision tree (DT), support vector machine (SVM), and gradient boosting decision tree (GBDT). Finally, an optimization model with the best sensitivity and specificity was selected for ENDOANGEL-LA and identifying independent factors associated with EGC. The framework of this study is illustrated in Figure 1. ENDOANGEL-LA was trained using Python 3.5 and the Keras library (v2.1.5) with Tensorflow 1.12.0 backend.

Prospective video test

ENDOANGEL-LA was tested with prospective videos to evaluate the performance of identifying EGC in clinical practice. We estimated the accuracy of ENDOANGEL-LA was 88.00% when diagnosing EGC in M-IEE videos and the accuracy of sole deep convolutional neural network (DCNN) was 72.00% based on a previous pilot study. The estimated sample size was 94 with a type I error rate of 0.05 and power of 0.80. All the videos of M-IEE were prospectively and consecutively collected from RHWU between Aug 18, 2020 and July 26, 2021.

The inclusion criteria were as follows: (1) age ≥18 years, (2) known gastric lesions needed to be further clarified under M-IEE, 3) signed informed consent. The exclusion criteria were as follows: (1) after gastrectomy; (2) lesions difficult to observe under M-IEE due to active bleeding, thick white coats, blurs, and mucus; (3) without pathological results; (4) pathologically confirmed advanced GC or lymphoma; (5) have participated in clinical trials of drugs and have been in the elution period of experimental drugs or control drugs; (6) previous history of allergy to anesthetic or spasmolytic.

According to pathological results, the unprocessed videos were edited into clips containing target lesions. When the frame was frozen by endoscopists during the examination, ENDOANGEL-LA were activated, extracted images features, analyzed the quantitative and deep learning feature indexes, and made the diagnosis of the frames. The results of every feature index and the final diagnosis were present on the screen for reference. (Figure S7 and Videos 1 and 2)

Comparing the performance of ENDOANGEL-LA and endoscopists

The randomly shuffled test set was used to compare the performance between ENDOANGEL-LA and the endoscopists, which include two experts (with more than ten years of experience of endoscopy), two seniors (with more than five years of experience of endoscopy), and four novices (with more than one year of experience of endoscopy). All the endoscopists were not enrolled in the annotation of the images and were blinded to patient information, endoscopy reports, pathological results, and diagnosis of ENDOANGEL-LA or other endoscopists. To explore the ENDOANGEL-LA assistance ability, after at least two weeks six endoscopists (one expert, one senior, and four novices) were asked to make a diagnostic decision again on the same test set with the assistance of ENAOANGEL-LA, and compared it with that of their independent performance.

Comparing the performance of ENDOANGEL-LA and sole DL models

To compare the performance between ENDOANGEL-LA and sole DL model in diagnosing EGC, ResNet-50 was used for constructing a sole DCNN model using the same training set. In addition, another sole DL model named ENDOANGEL-ME (magnifying endoscopy) was also used for the comparison, which was constructed from our previous studies using 4667 M-IEE images.[16]

Comparing the satisfaction level of ENDOANGEL-LA and sole DL models

Among the two sole DL models, the model with comparable diagnostic performance to ENDOANGEL-LA was selected for satisfaction evaluation. We estimated the satisfaction scores of ENDOANGEL-LA was 4.6 in M-IEE videos and the satisfaction scores of sole DCNN was 3.6 based on a previous pilot study. The estimated sample size of videos was 8 with a type I error rate of 0.05 and power of 0.80. Therefore, we used 100 prospective videos for satisfaction evaluation to fully meet the statistical requirement. Seventeen endoscopists as we could find, who did not participate in the criteria defining and the man-machine contest, were invited to review the 100 prospective videos mentioned above. Each case had an original video and two processed videos generated from ENDOANGEL-LA and the selected sole DL model based on the original one. Endoscopists were requested to finish an online questionnaire (Figure S8) after reviewing the videos. The Five-Point Likert Scale was used for assessing the extent of endoscopists' satisfaction for ENDOANGEL-LA and the selected sole DL model. Endoscopists also chosen which kind of AI systems they prefer to use (or not use any AI systems) in clinical practice.

Ethics

This study was approved by the Ethics Committee of RHWU and registered with trial number ChiCTR2000035116 in the WHO Registry Network's Primary Registries. For retrospective image data, informed consent was exempted by the institutional review boards.

Statistical analysis

The performance of ENDOANGEL-LA, sole DL models, and endoscopists was evaluated by metrics including accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and area under the curve (AUC). The optimal threshold of the receiver operating characteristic (ROC) curve was determined by Youden Index. The McNemar test was used to compare the accuracy, sensitivity, and specificity, and the Chi-square test was used to compare the PPV and NPV between ENDOANGEL-LA, sole DL models, and endoscopists. A Wilcoxon signed rank test was used to carry out the comparative analysis between ENDOANGEL-LA and sole DL models in satisfaction level of endoscopists. P-values < 0.05 were considered statistically significant. Interobserver agreement of endoscopists was evaluated using Cohen's kappa coefficient. All analyses were performed with SPSS (ver. 26.0; IBM, USA).

Role of the funding source

The funder had no role in study design, data collection, data analysis, data interpretation, writing of the report, or decision to submit the paper for publication. All authors had access to the raw datasets and accept responsibility for the decision to submit for publication.

Results

The analysis of feature indexes and performance of ENDOANGEL-LA in images

Among the seven ML models, RF showed the best performance in diagnosing EGC and was selected for the constriction of ENDOANGEL-LA. (The performance of seven ML models is shown in Figure 3 and Table S5) The feature indexes determined by RF are as follows: (1) density of MS, (2) eccentricity of MS equivalent centroid, (3) cyclization of MV, (4) spectral principal component information of gastric mucosal background color, (5) the arrangement of the M-IEE images, (6) the demarcation line of lesions, (7) the distribution of MV in the MV segmentation images, (8) the morphology of the lesions. And the corresponding weights of each feature index were 0.26, 0.07, 0.02, 0.04, 0.23, 0.28, 0.08, 0.02. (Figure 3) The AUC of ENDOANGEL-LA was 92.83% (95% confidence interval [CI]: [88.42%−96.24%]). In the test set, the ENDOANGEL-LA achieved a diagnostic accuracy of 88.76% (95%CI: [84.41%−92.01%]), a sensitivity of 86.39% (95%CI: [79.91%−91.01%]), a specificity of 91.67% (95%CI: [85.34%−95.41%]), a PPV of 92.70% (95%CI: [87.08%−95.99%]), a NPV of 84.62% (95%CI: [77.43%−89.8%]) The most false positives were intestinal metaplasia (IM) and chronic inflammation. The classification of errors is shown in Table S6, and the representative images of misdiagnosis are shown in Figure S9.

The performance of ENDOANGEL-LA in videos

A total of 123 patients undergoing M-IEE in RHWU were consecutively collected. 46 patients who met the

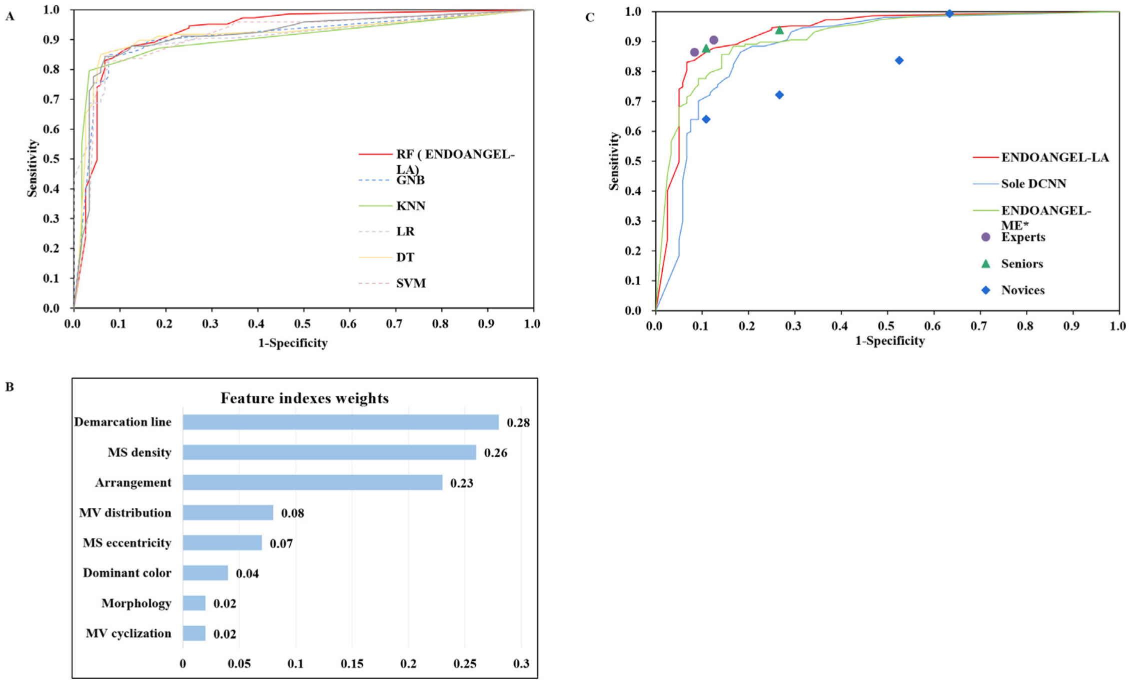

Figure 3. The corresponding weights of each feature index, the ROC curves of all ML models, ENDOANGEL-LA, sole DCNN, and ENDOANGEL-ME, and the performance of endoscopists. (A) The ROC curves of all ML models. (B) The corresponding weights of each feature index. (C) The ROC curves of ENDOANGEL-LA, sole DCNN, and ENDOANGEL-ME, and the performance of endoscopists.

MS: microsurfaces, MV: microvessels, ROC: receiver operating characteristic, ML: machine learning, RF: random forest, LA: logical anthropomorphic, GNB: Gaussian Naive Bayes, KNN: k-Nearest Neighbor, LR: logistic regression, DT: decision tree, SVM: support vector machine, GBDT: gradient boosting decision tree, DCNN: deep convolutional neural network, ME: magnifying endoscopy.

*He X, Wu L, Dong Z, Gong D, Jiang X, Zhang H, Ai Y, Tong Q, Lv P, Lu B, Wu Q, Yuan J, Xu M, Yu H. Real-time use of artificial intelligence for diagnosing early gastric cancer by magnifying image-enhanced endoscopy: a multicenter, diagnostic study (with videos). Gastrointest Endosc. 2022;95(4):671-678.e4.

exclusion criteria were excluded. Eventually, 77 patients including 25 EGC and 75 non-cancerous videos were included. The patient and lesion characteristics are shown in Table S7. ENDOANGEL-LA achieved an accuracy of 87.00% (95%CI: [79.02%−92.24%]), a sensitivity of 84.00% (95%CI: [65.35%−93.60%]), a specificity of 88.00% (95%CI: [78.74%−93.56%]), a PPV of 70.00% (95%CI: [52.12%−83.35%]), and a NPV of 94.29% (95%CI: [86.21%−97.76%]).

Comparison between ENDOANGEL-LA and endoscopists

Compared with novices in images, ENDOANGEL-LA had a better performance in accuracy (88.76% vs 71.63%, $p < 0.001$), sensitivity (86.39% vs 79.76%, $p = 0.002$), specificity (91.67% vs 61.67%, $p < 0.001$), PPV (92.70% vs 71.82%, $p < 0.001$), and NPV (84.62% vs 71.33%, $p < 0.001$) significantly. The specificity (81.25%, $p = 0.001$) and PPV (85.58%, $p = 0.006$) of seniors were significantly lower than those of ENDOANGEL-LA. In addition, the performance of ENDOANGEL-LA was comparable to that of the experts in images and that of endoscopists in videos. The comparison results are summarized in Figure 3 and Table 1. The details of the diagnosis performance of endoscopists are shown in Table S8. The inter-observer agreement between endoscopists is shown in Table S9. The most false positives of endoscopists were IM, chronic inflammation, and atrophy. (Table S6 and Figure S9) As shown in Table 1, the accuracy, sensitivity, specificity, PPV, and NPV of the novices with the assistance of ENAOANGEL-LA were significantly improved to 87.45% ($p < 0.001$), 85.03% ($p = 0.012$), 90.42% ($p < 0.001$), 91.58% ($p < 0.001$), and 83.14% ($p < 0.001$) respectively. (The details of the diagnosis performance of endoscopists with the assistance of ENAOANGEL-LA are shown in Table S8).

Comparison between ENDOANGEL-LA and sole DL models

In image test set, the AUC of ENDOANGEL-LA (92.83%, 95%CI: [89.42%−96.24%]) was better than that of sole DCNN (88.95%, 95%CI: [84.67%−93.22%]) and ENDOANGEL-ME (91.59%, 95%CI: [88.17%−95.01%]). The accuracy of ENDOANGEL-LA in diagnosing EGC was significantly higher than that of the sole DCNN (88.76% vs 82.77%, $p = 0.034$). The performance was comparable between ENDOANGEL-LA and ENDOANGEL-ME. In video test set, the accuracy (87.00% vs 68.00%, $p < 0.001$), specificity (88.00% vs 69.33%, $p = 0.001$) and PPV (70.00% vs 41.03%, $p = 0.007$) of ENDOANGEL-LA were significantly higher than those of sole DCNN. The accuracy (78.00%, $p < 0.001$) and PPV (53.85%, $p = 0.007$) of ENDOANGEL-ME were significantly lower than those of ENDOANGEL-LA. (Figure 3 and Table 1) The most false positives of sole DL models were IM, chronic inflammation, and atrophy. (Table S6 and Figure S9)

Comparison of satisfaction level between ENDOANGEL-LA and sole DL model

To evaluate and compare the satisfaction level of ENDOANGEL-LA and sole DL model, ENDOANGEL-ME, the diagnostic performance of which was comparable to that of ENDOANGEL-LA, was chosen in this test. The average satisfaction scores of endoscopists towards ENDOANGEL-LA were 4.76±0.42, which was significantly higher than that towards the ENDOANGEL-ME (3.76±0.64, $p = 0.001$). ENDOANGEL-LA (3.76±0.81) would affect endoscopists' judgment more than ENDOANGEL-ME (3.29±0.75, $p = 0.033$). More endoscopists prefer using ENDOANGEL-LA in clinical practice (Table 2) .

Discussion

In the study, we innovatively developed a feature extraction-based logical anthropomorphic diagnostic system named ENDOANGEL-LA for diagnosing EGC under M-IEE. ENDOANGEL-LA performed better than sole DL models and was comparable with experts, and gained better satisfaction from endoscopists. To our best knowledge, this is the first study developing a logical anthropomorphic AI system based on feature extracting to diagnose EGC with both good accuracy and explainability, and concretize abstract diagnostic theories.

At present, EGC is diagnosed under endoscopy relying on the experience summary of experts, which is in the form of descriptive and abstract theories.[12] However, each endoscopist has a different understanding of those abstract theories.[19] As a result, the consistency of EGC diagnosis between endoscopists is poor when the existing theories of EGC diagnosis are applied to clinical practice. AI has been widely applied in medical image-based disease determination and classification, and the development of AI systems for assisting the diagnosis of EGC has been an attractive research topic during the past decade.[20,21] However, existing AI models have always been questioned for their black-box nature.[22] Comments from experts and societies stated that AI systems without providing diagnostic logic may reduce the physicians' confidence and put patients at risk.[23–25] The existing AI models did not concretize abstract diagnostic theories to endoscopists and explain how the predictions were made, which remain fatal limitations for practical use.[23,26]

In this study, ENDOANGEL-LA achieved the concretization and explainability of diagnostic logic by simply and intuitively displaying the extracted feature indexes and diagnosis results on the screen as a diagnostic reference. We innovatively evaluate the features

		Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
Image test						
ENDOANGEL-LA		88.76% (84.41%–92.01%)	86.39% (79.91%–91.01%)	91.67% (85.34%–95.41%)	92.70% (87.08%–95.99%)	84.62% (77.43%–89.82%)
Sole DCNN		82.77% (77.78%–86.83%)*	82.31% (75.34%–87.63%)	83.33% (75.65%–88.94%)	85.82% (79.11%–90.63%)	79.37% (71.49%–85.52%)
ENDOANGEL-ME#		85.05% (80.24%–88.80%)	85.71% (79.15%–90.46%)	84.17% (76.59%–89.63%)	86.90% (80.44%–91.45%)	82.79% (75.12%–88.46%)
Without ENDOANGEL-LA	**Experts (*n* = 2)**	88.95% (86.01%–91.34%)	88.44% (84.28%–91.61%)	89.58% (85.07%–92.84%)	91.23% (87.37%–93.99%)	86.35% (81.53%–90.07%)
	Seniors (*n* = 2)	86.52% (83.36%–89.16%)	90.82% (86.97%–93.61%)	81.25% (75.83%–85.68%)*	85.58% (81.25%–89.05%)*	87.84% (82.89%–91.51%)
	Novices (*n* = 4)	71.63% (68.85%–74.25%)*	79.76% (76.32%–82.81%)*	61.67% (57.24%–65.91%)*	71.82% (68.25%–75.14%)*	71.33% (66.80%–75.47%)*
With ENDOANGEL-LA	**Expert (*n* = 1)**	89.89% (85.69%–92.96%)	87.76% (81.48%–92.12%)	92.50%(86.36%–96.00%)	93.48% (88.07%–96.53%)	86.05% (79.02%–90.99%)
	Senior (*n* = 1)	88.01% (83.56%–91.38%)	95.24% (90.50%–97.68%)	79.17% (71.06%–85.47%)	84.85% (78.56%–89.52%)	93.14% (86.51%–96.64%)
	Novices (*n* = 4)	87.45% (85.33%–89.30%)†	85.03% (81.92%–87.69%)†	90.42% (87.46%–92.74%)†	91.58% (88.95%–93.63%)†	83.14% (79.69%–86.11%)†
Video test						
ENDOANGEL-LA		87.00% (79.02%–92.24%)	84.00% (65.35%–93.60%)	88.00% (78.74%–93.56%)	70.00% (52.12%–83.35%)	94.29% (86.21%–97.76%)
Sole DCNN		68.00% (58.34%–76.33%)*	64.00% (44.52%–79.75%)	69.33% (58.17%–78.61%)*	41.03% (27.08%–56.59%)*	85.25% (74.28%–92.04%)
ENDOANGEL-ME#		78.00% (68.93%–85.00%)*	84.00% (65.35%–93.60%)	76.00% (65.22%–84.25%)	53.85% (38.57%–68.44%)*	93.44% (84.31%–97.42%)
Endoscopists		89.00% (81.37%–93.75%)	80.00% (60.87%–91.14%)	92.00% (83.63%–96.28%)	76.92% (57.95%–88.96%)	93.24% (85.13%–97.08%)

***Table 1*: The performance of ENDOANGEL-LA, sole DL models, and endoscopists.**

LA: logical anthropomorphic, DL: deep learning, CI: confidence interval, DCNN: deep convolutional neural network, ME: magnifying endoscopy, PPV: positive predictive value, NPV: negative predictive value.

* Significant difference between the target group and ENDOANGEL-LA ($p < 0.05$).

† Significant difference between the results of without ENDOANGEL-LA and with ENDOANGEL-LA ($p < 0.05$).

# He X, Wu L, Dong Z, Gong D, Jiang X, Zhang H, Ai Y, Tong Q, Lv P, Lu B, Wu Q, Yuan J, Xu M, Yu H. Real-time use of artificial intelligence for diagnosing early gastric cancer by magnifying image-enhanced endoscopy: a multicenter, diagnostic study (with videos). Gastrointest Endosc. 2022;95(4):671-678.e4.

	ENDOANGEL-LA	ENDOANGEL-ME[#]	P-value
Satisfaction scores	4.76±0.42	3.76±0.64	0.001*
The extent to which systems affect the judgment of endoscopists	3.76±0.81	3.29±0.75	0.033*
The number of endoscopists that tend to use one of the systems	16	1	–

Table 2: **Comparison between ENDOANGEL-LA and ENDOANGEL-ME in satisfaction assessment of endoscopists.**
LA: logical anthropomorphic, ME: magnifying endoscopy.

* Significant difference between ENDOANGEL-LA and ENDOANGEL-ME ($p < 0.05$).

[#] He X, Wu L, Dong Z, Gong D, Jiang X, Zhang H, Ai Y, Tong Q, Lv P, Lu B, Wu Q, Yuan J, Xu M, Yu H. Real-time use of artificial intelligence for diagnosing early gastric cancer by magnifying image-enhanced endoscopy: a multicenter, diagnostic study (with videos). Gastrointest Endosc. 2022;95(4):671-678.e4.

of EGC from abstract theories using quantitative algorithms and computer science and train the ML model to automatically screen-out independent features for the model.[10,17,27] Based on the independent features, the ML model assigned weights to the features automatically, from which the diagnostic value of different features can be explored for EGC. By deconstructing the decision-making process of ENDOANGEL-LA and exploring the relationship between different features in the future, it may be possible to learn some diagnostic rules of EGC that are difficult to be mastered by experience and provide a new basis for existing diagnostic theories, which may improve the accuracy and consistency of endoscopists.

According to the strong and detailed diagnostic reference given by ENDOANGEL-LA, the endoscopists can learn from the successful cases of model prediction, and improve the model by analyzing the reasons for the failure cases of the model prediction. Compared with ENDOANGEL-ME without explainability, endoscopists were more satisfied with ENDOANGEL-LA with explainability and had higher acceptance. They were more willing to use ENDOANGEL-LA in clinical practice. It can be seen that ENDOANGEL-LA with explainability was expected to improve human-machine collaboration and ensure the safety of the clinical application of AI.

Moreover, we evaluated the performance of the endoscopists with the assistance of ENAOANGEL-LA, and compared it with that of their independent performance. The comparison results showed that ENAOANGEL-LA could effectively improve the diagnostic performance of novices, which demonstrate the additional value of the ENAOANGEL-LA in the auxiliary diagnosis of EGC. In the error analysis, ENDOANGEL-LA did not misdiagnose any atrophy, but both endoscopists and ole DL models diagnosed atrophy as EGC. Therefore, with the assistance of ENDOANGEL-LA, inexperienced endoscopists may continuously improve the accuracy in diagnosing EGC.

In clinical practice, the endoscopic images are usually complex and diverse, and some features of the images may escape the naked eye.[28] Even among expert endoscopists, diagnosis often varies widely.[17] With the segmentation of MS and MV in the M-IEE images, ENDOANGEL-LA eliminated the influence of interference information on the endoscopists' observation and simplified the complex endoscopic image.[28] By extracting the vital diagnostic information and marking the most characteristic MV for reference, endoscopists can notice meaningful features of the images, which can enhance the endoscopists' confidence in diagnosis and may improve the accuracy of diagnosis.

Another advantage of this study is that the training sample size can be reduced. The DL algorithms always require large training samples (ranging from hundreds to millions) to learn and construct an appropriate model.[29,30] However, it is not easy to collect large data sets in clinical practice. In this study, the performance of ENDOANGEL-LA was better than sole DCNN using the same training set and comparable with ENDOANGEL-ME using 4667 M-IEE images.[16] The feature indexes related to EGC were determined in advance by experts, which may also present in non-cancerous lesions. Based on the given diagnosis logic, the model did not require a large number of EGC images to learn and summarize the difference between EGC and non-cancer.[31,32]

This study has several limitations. First, this study used clinical data collected from only one institution. We will further enhance the robustness of ENDOANGEL-LA by using multi-center data in the future. Second, this study only included M-IEE images with high definition. A clear view of MS and MV patterns is the prerequisite for quantifying the analysis. Nevertheless, high requirements for M-IEE images apparently indicate the qualified standard for M-IEE operation, which is conducive to the quality control of M-IEE and the detection of early cancer. Third, we did not conduct a clinical trial. However, the ENDOANGEL-LA was tested using consecutive and prospective M-IEE videos and performed well.

In conclusion, this study proposed a logical anthropomorphic AI system named ENDOANGEL-LA aimed at diagnosing EGC under M-IEE. ENDOANGEL-LA, which concretized abstract diagnostic theories and had good explainability, had a better performance than sole DL models. The performance of ENDOANGEL-LA was comparable to that of experts, and it had a promising role in future endoscopic scenarios.

Data sharing statement

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

Contributors

Honggang Yu and Lianlian Wu were responsible for conceiving and designing the study; Jia Li and Yijie Zhu were responsible for training and testing the models; Jia Li, Yijie Zhu, Zehua Dong, Xinqi He, Ming Xu, Liu Jun, Mengjiao Zhang, Xiao Tao, Hongliu Du, Di Chen, Li Huang, Renduo Shang, Lihui Zhang, Renquan Luo, Wei Zhou, Yunchao Deng, Xu Huang, Yanxia Li, Boru Chen, Rongrong Gong, Chenxia Zhan, and Xun Li were responsible for collecting and reviewing images; Jia Li and Yijie Zhu were responsible for collecting, collating and analyzing the data; Jia Li, Yijie Zhu, and Zehua Dong were responsible for writing the manuscript; Jia Li, Yijie Zhu, Zehua Dong, and Lianlian Wu were responsible for revising the manuscript; Honggang Yu was responsible for performing 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. Jia Li, Yijie Zhu, Zehua Dong, Lianlian Wu, and Yu Honggang have verified the data, all of them are independent of the company or investor. All authors contributed to critical revision of the manuscript.

Declaration of interests

We declare no competing interests.

Acknowledgement

This work was partly supported by a grant from the Hubei Province Major Science and Technology Innovation Project (2018-916-000-008) and the Fundamental Research Funds for the Central Universities (2042021kf0084).

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.eclinm.2022.101366.

References

1 Karimi P, Islami F, Anandasabapathy S, Freedman ND, Kamangar F. Gastric cancer: descriptive epidemiology, risk factors, screening, and prevention. *Cancer Epidemiol Biomark Prev*. 2014;23(5):700–713.

2 Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68(6):394–424.

3 Chen W, Zheng R, Baade PD, et al. Cancer statistics in China, 2015. *CA Cancer J Clin*. 2016;66(2):115–132.

4 Sung H, Ferlay J, Siegel RL, 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):209–249.

5 Van Cutsem E, Sagaert X, Topal B, Haustermans K, Prenen H. Gastric cancer. *Lancet*. 2016;388(10060):2654–2664.

6 Bisschops R, Areia M, Coron E, et al. Performance measures for upper gastrointestinal endoscopy: a European Society of Gastrointestinal Endoscopy (ESGE) quality improvement initiative. *Endoscopy*. 2016;48(9):843–864.

7 Pasechnikov V, Chukov S, Fedorov E, Kikuste I, Leja M. Gastric cancer: prevention, screening and early diagnosis. *World J Gastroenterol*. 2014;20(38):13842–13862.

8 Li WQ, Qin XX, Li ZX, et al. Beneficial effects of endoscopic screening on gastric cancer and its optimal screening interval: a population-based study. *Endoscopy*. 2021 preprint. https://doi.org/10.1055/a-1728-5673.

9 Pimenta-Melo AR, Monteiro-Soares M, Libânio D, Dinis-Ribeiro M. Missing rate for gastric cancer during upper gastrointestinal endoscopy: a systematic review and meta-analysis. *Eur J Gastroenterol Hepatol*. 2016;28(9):1041–1049.

10 Yao K. Magnifying endoscopy for the diagnosis of early gastric cancer: establishment of technique, diagnostic system, and scientific evidence from Japan. *Dig Endosc*. 2021 preprint. https://doi.org/10.1111/den.14178.

11 Chiu PWY, Uedo N, Singh R, et al. An Asian consensus on standards of diagnostic upper endoscopy for neoplasia. *Gut*. 2019;68(2):186–197.

12 Hu H, Gong L, Dong D, et al. Identifying early gastric cancer under magnifying narrow-band images with deep learning: a multicenter study. *Gastrointest Endosc*. 2021;93(6):1333–1341.

13 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(10):957–967.

14 Horiuchi Y, Hirasawa T, Ishizuka N, et al. Performance of a computer-aided diagnosis system in diagnosing early gastric cancer using magnifying endoscopy videos with narrow-band imaging (with videos). *Gastrointest Endosc*. 2020;92(4):856–865.

15 Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*. 2018;6:52138–52160.

16 He X, Wu L, Dong Z, et al. Real-time use of artificial intelligence for diagnosing early gastric cancer by magnifying image-enhanced endoscopy: a multicenter, diagnostic study (with videos). *Gastrointest Endosc*. 2022;95 (4):671–678.

17 Muto M, Yao K, Kaise M, et al. Magnifying endoscopy simple diagnostic algorithm for early gastric cancer (MESDA-G). *Dig Endosc*. 2016;28(4):379–393.

18 Liu X, Wang C, Bai J, Liao G, Zhao Y. Hue-texture-embedded region-based model for magnifying endoscopy with narrow-band imaging image segmentation based on visual features. *Comput Methods Programs Biomed*. 2017;145:53–66.

19 East JE, Vleugels JL, Roelandt P, et al. Advanced endoscopic imaging: European Society of Gastrointestinal Endoscopy (ESGE) technology review. *Endoscopy*. 2016;48(11):1029–1045.

20 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(10221):350–360.

21 Redlich R, Opel N, Grotegerd D, et al. Prediction of individual response to electroconvulsive therapy via machine learning on structural magnetic resonance imaging data. *JAMA Psychiatry*. 2016;73(6):557–564.

22 Price WN. Big data and black-box medical algorithms. *Sci Transl Med*. 2018;10(471):eaao5333.

23 Kundu S. AI in medicine must be explainable. *Nat Med*. 2021;27(8):1328.

24 Chou YL, Moreira C, Bruza P, Ouyang C, Jorge J. Counterfactuals and causability in explainable artificial intelligence: theory, algorithms, and applications. *Inf Fusion*. 2022;81:59–83.

25 Lee H, Yune S, Mansouri M, et al. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nat Biomed Eng*. 2019;3(3):173–182.

26 Hosny A, Parmar C, Coroller TP, et al. Deep learning for lung cancer prognostication: a retrospective multi-cohort radiomics study. *PLoS Med*. 2018;15:(11) e1002711.

27 Kaise M, Kato M, Urashima M, et al. Magnifying endoscopy combined with narrow-band imaging for differential diagnosis of superficial depressed gastric lesions. *Endoscopy*. 2009;41(4):310–315.

28 Ubaldi L, Valenti V, Borgese RF, et al. Strategies to develop radiomics and machine learning models for lung cancer stage and histology prediction using small data samples. *Phys Med*. 2021;90:13–22.

29 Koppe G, Meyer-Lindenberg A, Durstewitz D. Deep learning for small and big data in psychiatry. *Neuropsychopharmacology*. 2021;46(1):176–190.

30 Sun Q, Liu Y, Chua T, Schiele B. Meta-transfer learning for few-shot learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019403–412.

31 Peng Z, Zechao L, Junge Z, Yan L, Guo-Jun Q, Jinhui T. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019441–449.

32 Shi JY, Wang X, Ding GY, et al. Exploring prognostic indicators in the pathological images of hepatocellular carcinoma based on deep learning. *Gut*. 2021;70(5):951–961.