EBioMedicine 62 (2020) 103146

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.elsevier.com/locate/ebiom

Research paper

Development and validation of a real-time artificial intelligence-assisted system for detecting early gastric cancer: A multicentre retrospective diagnostic study

Dehua Tang[a,#], Lei Wang[a,#], Tingsheng Ling[a,#], Ying Lv[a,#], Muhan Ni[a], Qiang Zhan[b], Yiwei Fu[c], Duanming Zhuang[d], Huimin Guo[a], Xiaotan Dou[a], Wei Zhang[a], Guifang Xu[a,*], Xiaoping Zou[a,*]

[a] Department of Gastroenterology, Nanjing Drum Tower Hospital, Affiliated Drum Tower Hospital, Medical School of Nanjing University, 321 Zhongshan Road, Nanjing, Jiangsu 210008, China
[b] Department of Gastroenterology, Wuxi People's Hospital, Affiliated Wuxi People's Hospital with Nanjing Medical University, Wuxi, Jiangsu 214023, China
[c] Department of Gastroenterology, Taizhou People's Hospital, The Fifth Affiliate Hospital with Nantong University, Taizhou, Jiangsu 225300, China
[d] Department of Gastroenterology, Gaochun People's Hospital, Nanjing, Jiangsu 211300, China

ARTICLE INFO

Article History:
Received 21 July 2020
Revised 21 October 2020
Accepted 11 November 2020
Available online xxx

Keywords:
Artificial intelligence
Early gastric cancer
Convolutional neural network
Detection

ABSTRACT

Background: We aimed to develop and validate a real-time deep convolutional neural networks (DCNNs) system for detecting early gastric cancer (EGC).
Methods: All 45,240 endoscopic images from 1364 patients were divided into a training dataset (35823 images from 1085 patients) and a validation dataset (9417 images from 279 patients). Another 1514 images from three other hospitals were used as external validation. We compared the diagnostic performance of the DCNN system with endoscopists, and then evaluated the performance of endoscopists with or without referring to the system. Thereafter, we evaluated the diagnostic ability of the DCNN system in video streams. The accuracy, sensitivity, specificity, positive predictive value, negative predictive value and Cohen's kappa coefficient were measured to assess the detection performance.
Finding: The DCNN system showed good performance in EGC detection in validation datasets, with accuracy (85.1%–91.2%), sensitivity (85.9%–95.5%), specificity (81.7%–90.3%), and AUC (0.887–0.940). The DCNN system showed better diagnostic performance than endoscopists and improved the performance of endoscopists. The DCNN system was able to process oesophagogastroduodenoscopy (OGD) video streams to detect EGC lesions in real time.
Interpretation: We developed a real-time DCNN system for EGC detection with high accuracy and stability. Multicentre prospective validation is needed to acquire high-level evidence for its clinical application.
Funding: This work was supported by the National Natural Science Foundation of China (grant nos. 81672935 and 81871947), Jiangsu Clinical Medical Center of Digestive System Diseases and Gastrointestinal Cancer (grant no. YXZXB2016002), and Nanjing Science and Technology Development Foundation (grant no. 2017sb332019).

1. Introduction

Gastric cancer (GC) is one of the most common types of cancer and remains the third leading cause of cancer-related deaths worldwide [1]. Because the symptoms are minimal in the early stage, GC is diagnosed at an advanced stage in most patients, with a 5-year survival rate of ⟨ 30% [2]. However, if GC can be detected and diagnosed at an early stage, then curative resection will be possible, which could increase the 5-year survival rate to ⟩ 95% [2]. Therefore, early detection is one of the most effective strategies for reducing the mortality of GC.

Endoscopy with white light imaging (WLI) is considered a standard modality for the detection of early GC (EGC) worldwide [3]. However, EGC often appears as delicate changes in the mucosa, making the overall sensitivity of WLI in detecting EGC not entirely satisfactory (40%–60%) [4]. Several studies have indicated that magnifying endoscopy in combination with image-enhanced endoscopy techniques, such as narrow-band imaging (NBI), auto-fluorescence imaging, and blue laser imaging, can remarkably improve the performance in detecting EGC [4-6]. However, these advanced devices,

* Corresponding authors.
E-mail addresses: 13852293376@163.com (G. Xu), zouxp@nju.edu.cn (X. Zou).
# These authors equally contributed to this project.

https://doi.org/10.1016/j.ebiom.2020.103146

Research in Context

Evidence before this study

Gastric cancer (GC) is mostly diagnosed at an advanced stage, with a 5-year survival rate of ⟨ 30%. But if GC can be diagnosed and then curatively resected at an early stage, which could increase the 5-year survival rate to ⟩ 95%. However, detection of early gastric cancer is a big challenge. White light imaging (WLI) is the most popular modality for detection of early gastric cancer, but the overall sensitivity of WLI is relatively low (40%−60%). Although other image-enhanced endoscopy techniques can improve the diagnostic ability of EGC, these advanced devices, together with experienced endoscopists, are not always available. Therefore, novel practical tools in detecting EGC lesions are needed.

Added value of this study

By using a total 45,240 endoscopic images from 1364 patients, this study developed and validated a real-time DCNN system for EGC detection. Another 1514 images from three other hospitals were used as external validation. The DCNN system showed good performance in EGC detection in different validation datasets. The DCNN system showed better diagnostic ability and stability in EGC detection than expert or trainee endoscopists. Moreover, the DCNN system was able to process oesophagogastroduodenoscopy (OGD) video streams to detect EGC lesions in real time. We also developed a website to provide free access to our DCNN system (http://112.74.182.39), with the an open-access database containing 300 cancerous lesions and 300 non-cancerous controls available upon reasonable request on the website.

Implications of all the available evidence

A real-time DCNN system for EGC detection with high accuracy and stability is developed and validated here, showing great potential in assisting endoscopists to detect EGC. Multicentre prospective validation is needed to acquire high-level evidence for its clinical application in EGC detection.

together with experienced endoscopists, are not always available, especially in rural or undeveloped areas. Whereas the detection rate of EGC is about 75% of all GCs in Japan, the detection rate in China is only 5%−20% [7, 8]. Therefore, the development of practical tools that can assist endoscopists in detecting EGC lesions is of great value.

In recent years, artificial intelligence (AI) systems based on deep convolutional neural network (DCNN) algorithms have achieved unprecedented success[9, 10]. In the medical field, AI systems have shown vast superiority in detecting skin cancers, diagnosing diabetic retinopathy, and improving the quality of oesophagogastroduodenoscopy (OGD) [11-13]. Several preliminary studies have applied AI to the detection of GC; however, limitations such as low efficiency [14], selection bias in datasets [15, 16], and applicability only for static images [14, 15], have compromised the clinical value.

We aimed to develop an efficient AI system based on DCNNs to detect EGC under WLI in real time.

2. Methods

2.1. Study design and participants

This study was performed at four institutions in China: Endoscopic Center of Nanjing University Medical School Affiliated Drum Tower Hospital (NJDTH), Endoscopic Center of Wuxi People's Hospital (WXPH), Endoscopic Center of Taizhou People's Hospital (TZPH), and Endoscopic Center of Gaochun People's Hospital (GCPH). A total of 1568 patients who underwent endoscopic submucosal dissection (ESD) according to associated guidelines [17] between January 2016 and January 2019 were retrospectively included in this study. Among these patients, 1508 were contributed by NJDTH and the remaining 60 were provided by the three other hospitals. The inclusion criteria were as follows: a diagnosis of EGC; ESD treatment; histologically proven malignancy; and endoscopic examination before ESD at NJDTH, WXPH, TZPH, or GCPH. The exclusion criteria were as follows: history of chemotherapy or radiation to the stomach, lesions adjacent to the ulcer or ulcer scar, gastric stump cancer, and multiple synchronous cancerous lesions.

2.2. Data preparation

A total of 80,791 endoscopic images from 1482 patients were retrospectively obtained from the imaging database of the NJDTH endoscopic center. Five experienced endoscopists from NJDTH (each of whom had > 5 years of experience and had performed at least 5000 OGD examinations) assessed the quality of all images. A total of 35,551 endoscopic images of NBI, dye-stained imaging, ESD operation, or poor quality (e.g., less insufflation of air, halation, defocus, blurs, bubbles, sliding, fuzzy, bleeding) were excluded from the study. The remaining 45,240 endoscopic images from 1364 patients were used for the development and temporal validation of the DCNN system. The temporal validation dataset was independent of the training dataset. Another retrospective dataset including 26 OGD videos with EGC lesions, which were independent of all the 45,240 static images, was used to assess the performance of the DCNN system in real time. To conduct external validation, we selected 406 images of 20 patients from WXPH, 556 images of 20 patients from TZPH, and 552 images of 20 patients from GCPH. All endoscopic images and videos were recorded using Olympus endoscopes (GIF-H260, GIF-H260Z, GIF-HQ290, GIF-H290Z; Olympus Medical Systems, Tokyo, Japan) with video processors (EVIS LUCERA CV260/CLV260SL, EVIS LUCERA ELITE CV290/CLV290SL, Olympus Medical Systems). All images were anonymised before inclusion to protect the privacy of the patients.

Specifically, the training and validation datasets were as follows: 1) The training dataset included 35,823 images of 1085 patients from NJDTH between January 2016 and October 2018 (among these images, 26,172 contained cancerous lesions). 2) The temporal validation dataset included 9417 images of 279 patients from NJDTH between November 2018 and January 2019 (among these images, 4153 contained cancerous lesions). 3) The external validation datasets included 406 images of 20 patients from WXPH (203 images contained malignant lesions), 556 images of 20 patients from TZPH (228 images contained malignant lesions), and 552 images of 20 patients from GCPH (226 images included cancerous lesions), and all of these images were obtained and filed between June 2019 and October 2019. 4) The video dataset included 26 videos of 26 patients from NJDTH between November 2019 and December 2019. 5) The testing dataset included 300 cancerous images and 300 control images (no malignant lesions in the images) randomly selected from the temporal validation dataset to compare the performance of the DCNN system and endoscopists. The control images contained several types of non-cancerous images, including chronic non-atrophic gastritis, chronic atrophic gastritis, and erosion (Table S1). The sample distribution is shown in Fig. 1.

Two board-certified pathologists determined the pathologic diagnosis of EGC using haematoxylin- and eosin-stained tissue slides, according to the World Health Organization (WHO) Classification of Tumours 5th edition. The same five experienced endoscopists studied the guidelines of the European Society of Gastrointestinal Endoscopy and Japanese Gastric Cancer Association. All selected images were

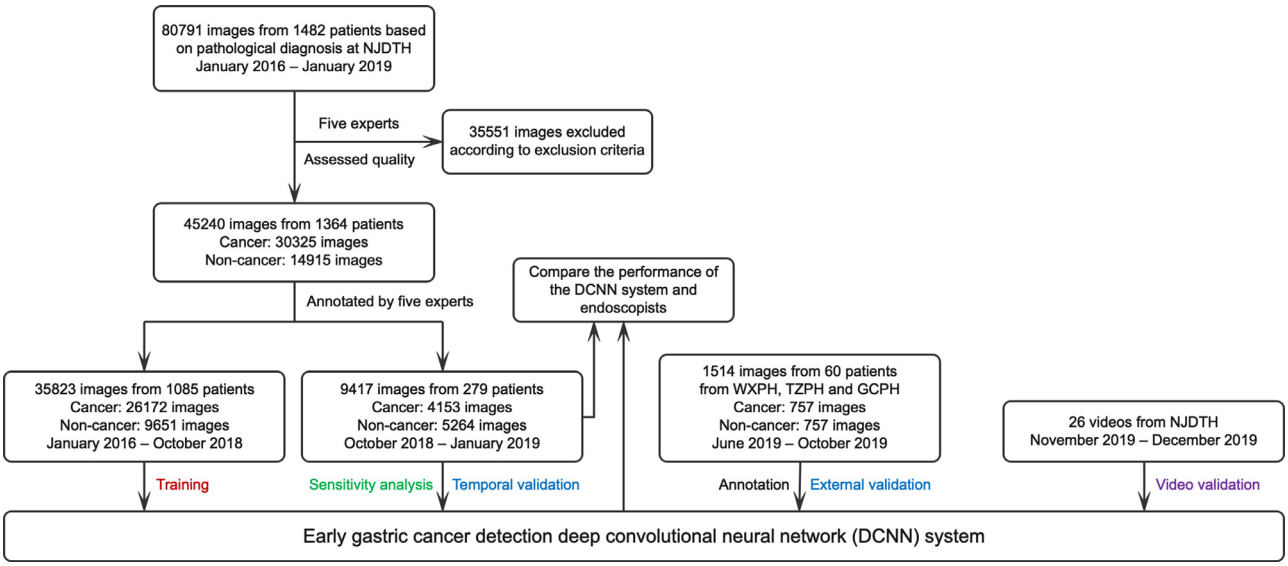

Fig. 1. Workflow for the development and validation of the DCNN system for diagnosing EGC. DCNN: Deep convolutional neural networks; EGC: Early gastric cancer.

categorised into the cancer (31,082 images) and non-cancer (15,672 images) groups. Thereafter, images from the cancer group were independently annotated by experienced endoscopists based on the pathological diagnosis. Specifically, the endoscopists were asked to outline the boundary of the actual lesion areas within the images. Thereafter, the circumscribed rectangles of the outlined regions were generated by the computer as annotation boxes. Thus, the size of the annotation boxes was precisely controlled. To avoid individual bias, the annotations and marks in the images were finalised only when more than four endoscopists have reached a consensus.

2.3. Training of the DCNN system

During the training process for the DCNN system, the parameters of the neurones in the network were initially set to random values. For each annotated image, the location of a lesion computed by the system was compared with the annotated areas. The parameters of this mathematical function were then slightly modified to decrease the error in the same image. The same process was repeated multiple times for every image in the training set.

The architecture of the deep network mainly comprises two parts: the backbone structure to extract features from the image and the detection and decision layers to detect the location of the lesions. The backbone structure employed in this model was the Darknet-53 model, which contains 53 layers of neurones. This architecture consists of a sequence of non-linear processing modules. Each module consists of one or more convolutional layers with batch normalization and Leaky ReLU non-linearity activation functions. Mainly, the modules of residual networks are employed in the architectures to increase the depth of the feature-extracting network. The decision layers predict four coordinates for each bounding box based on the features extracted from the convolutional layers. Three different scales of predictors were employed to detect large, medium, and small objects (Fig. 2). In the meanwhile, the decision layers will output the confidence of the content in the bounding boxes and classify the images into certain classes according to the cut-off value. The confidence of the results contains two parts, the Intersection-over-Union (IoU) to evaluate the overlap of predicted bounding boxes and the ground truth bounding boxes, which is defined as:

$$IoU = \frac{area(predicted\ bounding\ boxes) \cap area(ground\ truth\ bounding\ boxes)}{area(predicted\ bounding\ boxes) \cup area(ground\ truth\ bounding\ boxes)}$$

and the probability of classification (Pr(object)), which is the classification probability of object in the predicted bounding box. The total confidence in this detection task is defined as: $Confidence = Pr(object) \times IoU$.

2.4. Validation and testing of the DCNN system and comparison with endoscopists

First, we evaluated the performance of our DCNN system in the detection of EGC in patients using the independent temporal validation dataset. Second, we assessed the robustness of our DCNN system using the three external validation datasets from WXPH, TZPH, and GCPH. Third, we evaluated performance of the DCNN system in subgroups of EGC lesions with the temporal validation dataset. We divided the temporal validation dataset into three datasets, including intraepithelial lesions dataset, intramucosal lesions dataset and submucosal lesions dataset according to the cases. Then, we analyzed the diagnostic performance of the DCNN system in the three types of early gastric cancer lesions. Fourth, we compared the performance of the DCNN system and endoscopists using the testing dataset. Endoscopists from the four institutions were assigned to two groups based on level of expertise: 6 experts (minimum of 10-year experience with 10,000 OGD examinations) and 10 trainees (2-year experience with 2000 OGD examinations). These endoscopists were not involved in the selection and annotation of the image datasets, and were masked to the clinical characteristics, endoscopic manifestations, and pathological results of all patients. The testing images were all mixed in scrambled order and assessed by the endoscopists. Fifth, to assess the stability of the DCNN system and endoscopists, the same group of testing images was scrambled and assigned to the DCNN system and endoscopists for re-testing 3 days later. Sixth, we tested the performance of our DCNN system using a video dataset. The flowchart of this study is shown in Fig. 1. We also developed a website to provide free access to our DCNN system (http://112.74.182.39) (Fig. S4), with an open-access database containing 300 cancerous lesions and 300 non-cancerous controls available upon reasonable request on the website.

2.5. Outcomes

The accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of the diagnosis were the primary outcomes. Accuracy is defined as the system identifies a predicted box contains a cancerous lesion when its confidence value output by the

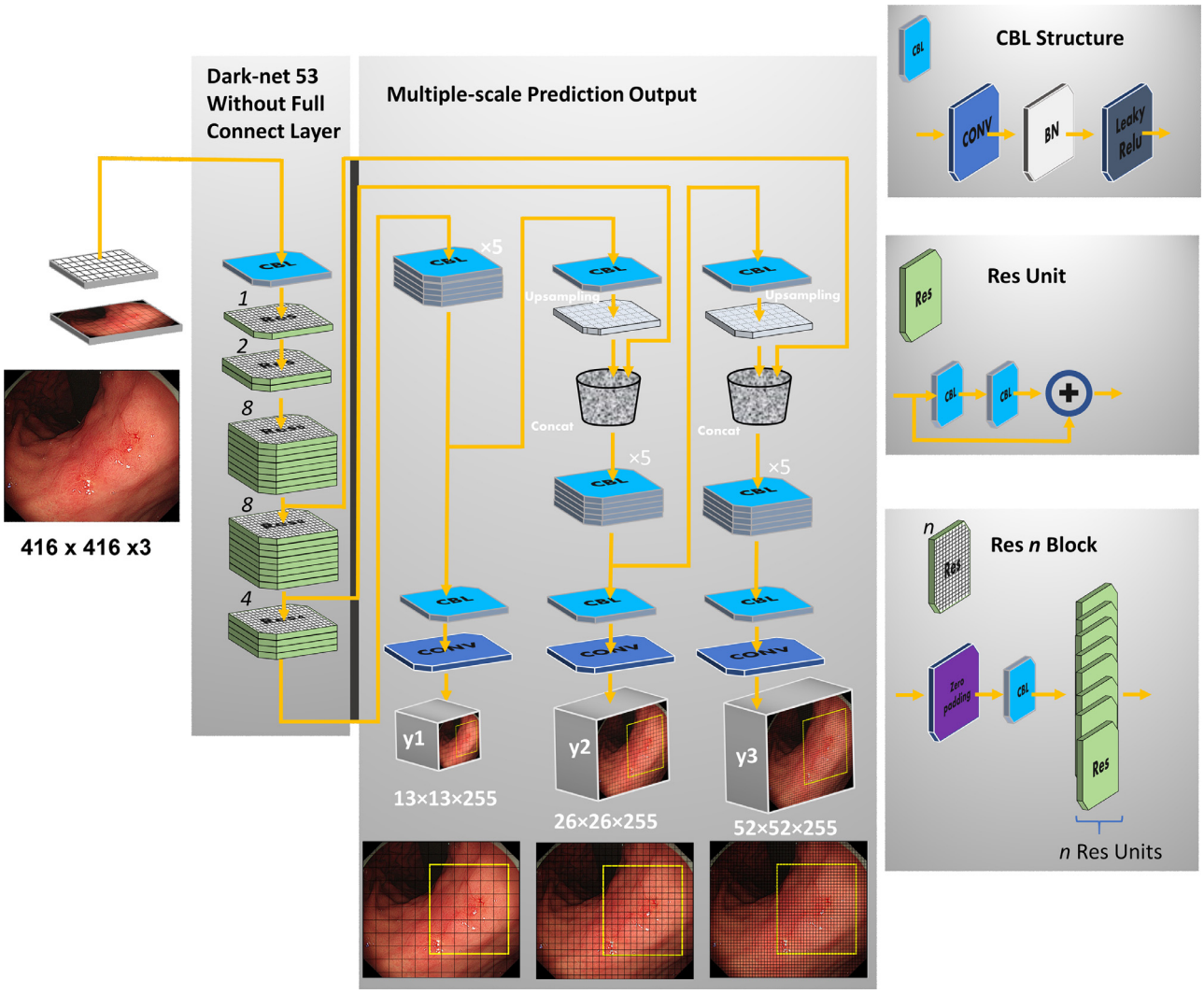

Fig. 2. Architecture and workflow of the DCNN system. DCNN: Deep convolutional neural networks. .

DCNN is bigger than a given cut-off value. In this study, the cut-off value is defined as the value for which the point on the Receiver operating characteristic (ROC) curve has the minimum distance to the upper left corner (where sensitivity=1 and specificity=1). Accuracy = true predictions/total number of cases, Sensitivity = true positive/total number of positive cases, Specificity = true negative/total number of negative cases, PPV = true positive/(true positive + false positive), NPV = true negative/(true negative + false negative).

2.6. Statistical analysis

A two-sided McNemar test was used to compare the differences in accuracy, sensitivity, specificity, PPV, and NPV of endoscopists before and after conferring to the results of DCNN. Receiver operating characteristic (ROC) curve analysis was employed, and the area under the ROC curve (AUC) was calculated to evaluate the diagnostic ability of the DCNN system in the temporal validation and external validation datasets. Inter-observer and intra-observer agreement of the endoscopists and DCNN were calculated using Cohen's kappa coefficient. Statistical analysis was performed using SPSS (version 26.0; IBM Inc., Armonk, NY, USA) or R software (version 3.6.3).

2.7. Ethics

The study design was reviewed and approved by the Medical Ethics Committee of Nanjing University Medical Affiliated Drum Tower Hospital (approval no. 2020−026−01). The study was registered in the WHO Registry Network's Primary Registries (ChiCTR2000031058). Informed consent was not required from patients whose images were retrospectively obtained from the image databases at each hospital involved in this study.

2.8. Role of the funding source

The funders had no role in study design, data collection, data analyses, interpretation, or writing of report. The corresponding authors had full access to all the data in the study.

3. Results

3.1. Performance of the DCNN system in detecting EGC lesions

The clinical characteristics of the patients enrolled in this study are shown in Table 1. The diagnostic ability of our DCNN system in detecting EGC lesions was evaluated using four independent validation datasets (Table 2). In the NJDTH validation dataset, the diagnostic accuracy was 87.8% (Fig. S1a). In the three external validation datasets, the accuracies were 88.7% for WXPH, 91.2% for TZPH, and 85.1% for GCPH (Fig. S1b and S1c and S1d). The sensitivity and NPV of our DCNN system were > 85% for all validation datasets. The specificity of the DCNN system ranged from 81.7% to 90.3%, and the PPV ranged

Table 1
Clinical characteristics of training and validation datasets.

Characteristics	Training dataset (NJDTH, 1085 cases) January 2016–October 2018	Temporal validation dataset (NJDTH, 279 cases) November 2018–January 2019	External validation datasets June 2019–October 2019			Video dataset (NJDTH, 26 cases) November 2019–December 2019
			WXPH (20 cases)	TZPH (20 cases)	GCPH (20 cases)	
Sex (male/female)	808 / 277	191 / 88	13 / 7	12 / 8	15 / 5	16 / 10
Age (years), mean (range)	63.4 (27–90)	64.0 (34–86)	66.3 (51–78)	61.8 (47–73)	62.6 (54–79)	62.6 (39–77)
Size (cm), mean (range)	2.1 (0.2–4.4)	1.9 (0.3–3.9)	1.9 (0.6–3.5)	1.7 (0.5–3.6)	1.4 (0.5–2.6)	1.7 (0.4–3.5)
Location (gastro-oesophageal junction / gastric fundus / gastric body / angulus / antrum)	340 / 8 / 194 / 164 / 379	92 / 4 / 41 / 40 / 102	12 / 0 / 0 / 3 / 5	8 / 0 / 2 / 4 / 6	7 / 0 / 0 / 4 / 9	8 / 0 / 4 / 6 / 8
Macroscopic type (I / IIa / IIb / IIc / IIa + IIc / IIc + IIa / IIb + IIc / III)	275 / 183 / 48 / 363 / 150 / 27 / 10 / 29	32 / 68 / 16 / 107 / 45 / 5 / 3 / 3	0 / 4 / 1 / 8 / 4 / 3 / 0 / 0	0 / 6 / 1 / 6 / 3 / 3 / 1 / 0	2 / 4 / 0 / 5 / 4 / 3 / 1 / 1	0 / 3 / 0 / 17 / 5 / 0 / 1 / 0
Degree of differentiation (differentiated / undifferentiated / mixed)	1002 / 14 / 69	256 / 2 / 21	19 / 0 / 1	17 / 1 / 2	19 / 0 / 1	24 / 0 / 2
Invasion depth (LGD + HGD / M / SM)	471 / 368 / 246	99 / 135 / 45	4 / 15 / 1	3 / 17 / 0	6 / 12 / 2	3 / 22 / 1
Atrophy (with / without)	799 / 286	173 / 106	17 / 3	13 / 7	16 / 4	20 / 6

LGD: Low grade dysplasia; HGD: High grad dysplasia; M: Mucosal gastric cancer; SM: Submucosal gastric cancer.

Table 2
. Performance of the DCNN system in validation datasets.

	NJDTH validation Internal validation	External validation WXPH	TZPH	GCPH
Accuracy (95% CI)	87.8 (87.1–88.5)	88.7 (85.2–91.4)	91.2 (88.5–93.3)	85.1 (81.9–87.9)
Sensitivity (95% CI)	95.5 (94.8–96.1)	91.1 (86.1–94.5)	92.1 (88.1–94.9)	85.9 (81.0–89.7)
Specificity (95% CI)	81.7 (80.7–82.8)	86.2(80.5–90.5)	90.3 (86.0–93.4)	84.4 (79.5–88.4)
Positive predictive value (95% CI)	80.5 (79.4–81.6)	86.9 (81.4–90.9)	90.5 (86.3–93.5)	84.6 (79.8–88.6)
Negative predictive value (95% CI)	95.9 (95.2–96.4)	90.7 (85.4–94.2)	91.9 (87.9–94.8)	85.7 (80.8–89.5)
AUC	0.940	0.906	0.925	0.887

from 80.5% to 90.5%. The specificity and PPV in the NJDTH validation dataset were the lowest among all the validation datasets. Notably, the predictive area of EGC lesions with the DCNN system was in high accordance with the positive pathological tissues (Fig. 3a) and annotation by experts (Fig. 3b). High AUC values (0.887–0.940) indicated an excellent diagnostic performance of the DCNN system in the four validation datasets (Fig. 4). Sensitivity analysis showed that the DCNN system achieved comparable performance in three subgroups of EGC including intraepithelial lesions, intramucosal lesions and submucosal lesions (AUC: 0.938 vs 0.946 vs 0.937, Table S2 and Fig. S2 and Fig. S3).

3.2. Comparison between the DCNN system and endoscopists

We compared the diagnostic performance of the DCNN system and endoscopists using a testing dataset. As shown in Table 3, the accuracy of the DCNN system (95.3%) was higher than that of expert (87.3%) and trainee (73.6%) endoscopists. Moreover, the sensitivity, specificity, PPV, and NPV of the DCNN system were all superior to those of both groups of endoscopists. Although endoscopists from the expert and trainee groups achieved rather comparable specificity and PPV, the sensitivity and NPV were relatively higher in the expert group. We then combined and analysed the diagnostic ability of the DCNN system with that of endoscopists. The results showed that in the expert group, the diagnostic accuracy (87.3% (95% confidence interval [CI] 85.2%–89.3%) vs. 94.3% (95% CI 91.0%–97.5%), McNemar test, P = 0.002), sensitivity (82.7% (95% CI 75.5%–89.9%) vs. 97.4% (95% CI 95.0%–99.8%), McNemar test, P = 0.005), and NPV (85.4% (95% CI 80.1%–90.7%) vs. 97.9% (95% CI 96.3%–99.4%), McNemar test, P = 0.002) significantly increased after combining the results of the DCNN system. The specificity (91.9% (95% CI 87.2%–96.6%) vs. 91.1% (95% CI 83.1%–99.1%), McNemar test, P = 0.553) and PPV (92.1% (95% CI 88.4%–95.7%) vs. 92.1% (95% CI 85.6%–98.5%), McNemar test, P = 1.000) were comparable regardless of the assistance of the DCNN system. In the trainee group, the accuracy (73.6% (95% CI 71.0%–76.3%) vs. 96.2% (95% CI 95.8%–96.7%), McNemar test, $P < 0.001$), sensitivity (50.2% (95% CI 44.1%–56.4%) vs. 94.7% (95% CI 93.9%–95.6%), McNemar test, $P < 0.001$), specificity (97.1% (95% CI 95.6%–98.5%) vs. 97.7% (95% CI 96.8%–98.6%), McNemar test, P = 0.914), PPV (95.1% (95% CI 93.2%–96.9%) vs. 97.7% (95% CI 96.8%–98.5%), McNemar test, P = 0.039), and NPV (66.7% (95% CI 63.9%–69.5%) vs. 94.9% (95% CI 94.1%–95.7%), McNemar test, $P < 0.001$) all remarkably increased when combined with the predictive results of the DCNN system. Thereafter, we investigated the stability of the DCNN system and endoscopists. The DCNN system was stable at all circumstances; however, the performance of endoscopists showed fluctuations. Our results showed that expert endoscopists achieved much more substantial intra-observer agreement (κ: 0.727–0.802) than trainee endoscopists (κ: 0.355–0.744) (Table 4). The inter-observer agreement of experts was also higher than that of trainees (Table S3 and Table S4).

3.3. OGD videos with EGC lesions detected with the DCNN system

The trainee endoscopists required approximately 6.84 s per image to make a diagnosis, whereas the experts required 6.13 s per image (Table S5). The DCNN system needed only 15 ms to diagnose a single image, indicating that the system can diagnose > 60 images per 1 s in real time. On the basis of the extremely fast diagnostic speed, we tested the performance of the DCNN system in diagnosing EGC in a video dataset (Video 1 and Video 2). The DCNN system detected 23 lesions in 26 OGD videos. The sensitivity of the DCNN system for detecting lesions in OGD videos was 88.5% (95% confidence interval [CI]: 71.0%–96.0%). On the basis of the excellent performance of the DCNN system, we developed an AI-based diagnostic platform to assist in the detection of EGC in routine OGD examinations. We also

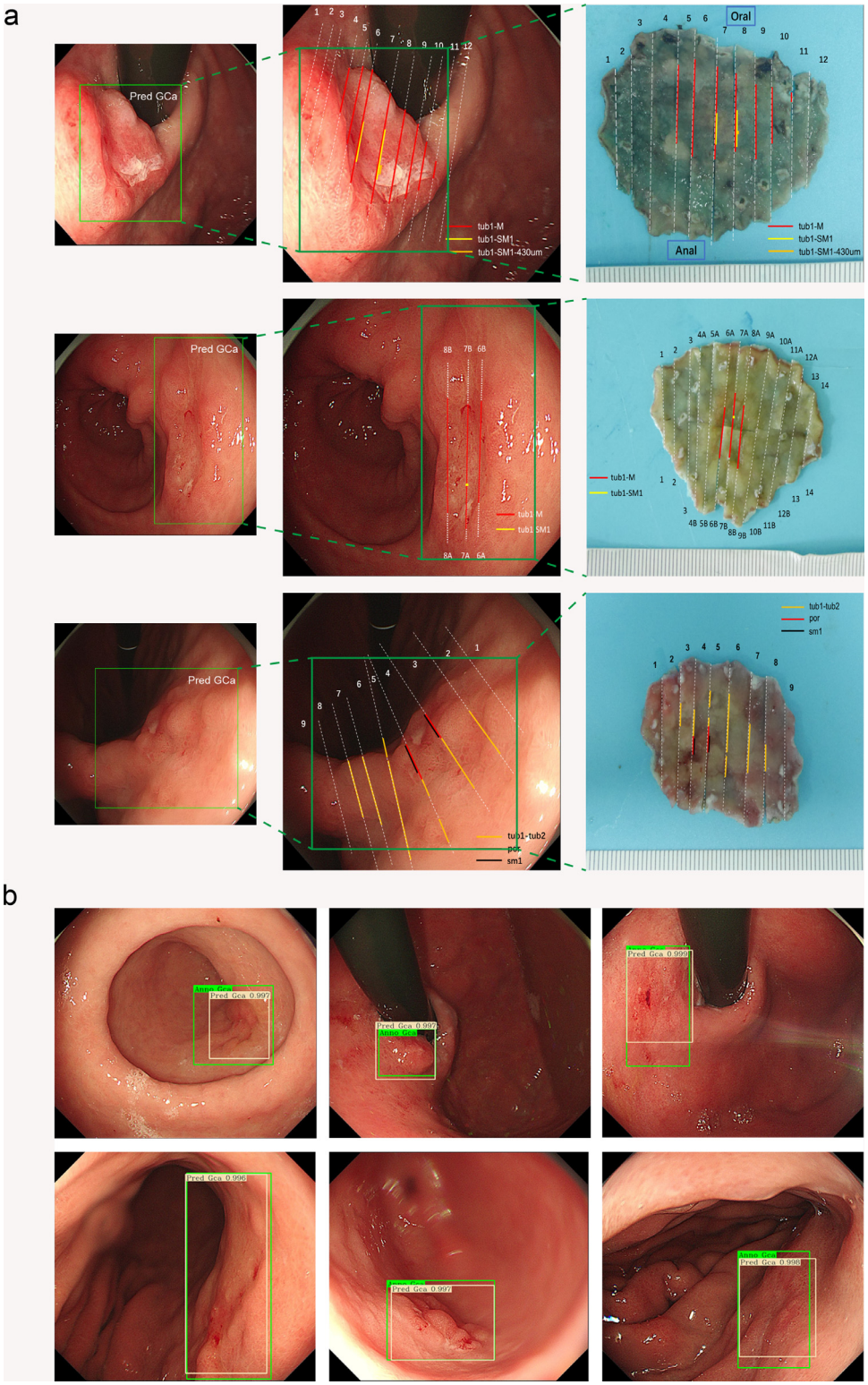

Fig. 3. (a) Predictive results of the DCNN system and corresponding positive pathological tissues. (b) Predictive results of the DCNN system and corresponding annotations of experts. DCNN: Deep convolutional neural networks. .

developed a website to provide free access to our DCNN system (http://112.74.182.39) (Fig. S4). We also made an open-access image database (the testing dataset) containing 300 cancerous lesions and 300 non-cancerous controls available on the website, which might be a useful resource for researchers in the field of AI-assisted medical imaging.

4. Discussion

In this study, we developed an AI system based on DCNNs to assist endoscopists in detecting EGC during OGD. The DCNN system demonstrated good diagnostic ability in independent validation datasets, with satisfactory accuracy (85.1%–91.2%), sensitivity (85.9%–95.5%),

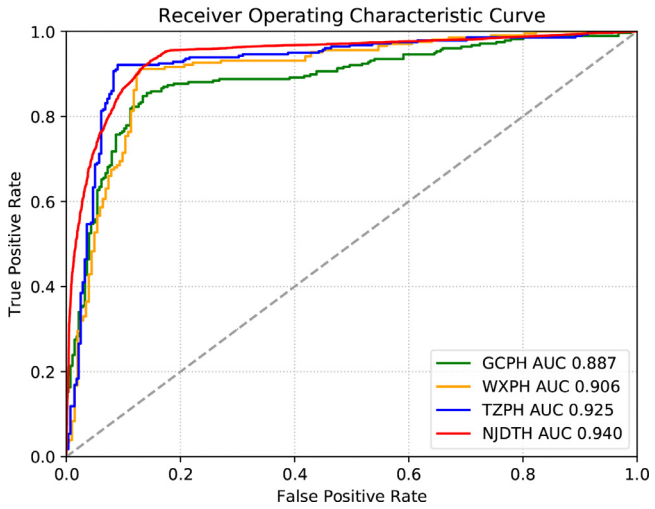

Fig. 4. Receiver operating characteristic curves illustrating the ability of the DCNN system to diagnose EGC. Sample size: 4153 cancer images and 5264 non-cancer images in NJDTH; 203 cancer images and 203 non-cancer images in WXPH; 228 cancer images and 228 non-cancer images in TZPH; 226 cancer images and 226 non-cancer images in GCPH. NJDTH: Nanjing University Medical School Affiliated Drum Tower Hospital; WXPH: Wuxi People's Hospital; TZPH: Taizhou People's Hospital; GCPH: Gaochun People's Hospital; DCNN: Deep convolutional neural networks; EGC: Early gastric cancer.

Table 4
Intra-observer agreement of the testing dataset.

Expert / trainee	κ
Expert 1	0.802
Expert 2	0.765
Expert 3	0.778
Expert 4	0.727
Expert 5	0.769
Expert 6	0.802
Trainee 1	0.552
Trainee 2	0.534
Trainee 3	0.535
Trainee 4	0.570
Trainee 5	0.419
Trainee 6	0.634
Trainee 7	0.355
Trainee 8	0.744
Trainee 9	0.662
Trainee 10	0.672

NPV (85.7%–95.9%), and AUC value (0.887–0.940). Sensitivity analysis showed that the DCNN system achieved comparable performance in three subgroups of EGC lesions including intraepithelial lesions, intramucosal lesions and submucosal lesions. The performance of the DCNN system in EGC detection was much better than that of endoscopists. The diagnostic performance of trainee endoscopists when combined with the DCNN system became comparable to that of expert endoscopists. Moreover, the DCNN system was able to process OGD videos to detect EGC lesions in real time owing to its extremely fast diagnostic speed (15 ms per image). To our knowledge, our DCNN system is the most efficient AI-based system for detecting EGC lesions worldwide.

Endoscopy with targeted biopsy is the gold standard method for diagnosing EGC [18]. However, as EGC lesions often appear as subtle mucosal changes under conventional WLI, the successful detection of these lesions largely depends on the skills and experience of endoscopists [19]. Previous studies have validated the diagnostic efficacy of WLI in detecting EGC, with rather unsatisfactory sensitivity (48%, 95% CI: 39%–57%) and specificity (67%, 95% CI: 62%–71%) [5]. Recently, several advanced technologies such as magnifying endoscopy, NBI, auto-fluorescence imaging, and blue laser imaging have been applied for the detection of EGC [4, 6, 20]. In several studies, the sensitivity (83%, 95% CI: 79%–87%) and specificity (96%, 95% CI: 95%–97%) of magnifying endoscopy combined with NBI in the diagnosis of EGC were superior to those of conventional WLI [5]. However, NBI has insufficient brightness, making it unsuitable for use in routine screening. Moreover, achieving sufficient sensitivity with these high technologies often depends on the expertise of endoscopists [21]. Our DCNN system was much more friendly to endoscopists, with nearly no requirements for experience and training. With the assistance of the DCNN system, the performance of trainee endoscopists in EGC detection significantly improved, with a sensitivity from 82.7% to 94.7%. Notably, with the assistance of the DCNN system, the performance of trainee endoscopists became comparable to that of experts (sensitivity: 94.7% vs. 97.4%). These results indicate that the DCNN system has great potential for improving the detection rate of EGC, especially for endoscopists lacking extensive experience and training in developing regions.

Several prior studies have reported the application of DCNNs in assisting in the diagnosis of GC. Toshiaki et al. developed a convolutional neural network (CNN)-based GC diagnostic system [14]. The sensitivity of this CNN system was 92.2%, but the PPV was only 30.6%. PPV is used to quantify the detection efficiency. The low PPV indicated that the cancerous lesions detected by this system included excessive false-positive judgements, which would increase the risk of bleeding and pose a heavy burden on endoscopists and pathologists by requiring them to conduct more biopsies and pathological diagnoses, respectively. We have given attention to both sensitivity and PPV in our DCNN system, and achieved an elegant balance with satisfactory sensitivity and acceptable PPV. This could enable the system to detect more potential EGC lesions and reduce unnecessary biopsies at the same time. Accordingly, the prediction of our system showed remarkable consistency with positive pathological tissues and annotation by experts. Hong et al. reported a lesion-based CNN for EGC detection with a sensitivity of 91.0% and an AUC value of 0.981 [15]. However, the validation dataset was a small subset randomly selected from the whole collected images, which indicated that several images from one patient might exist in both the training and validation datasets. This would, in turn, lead to overfitting. Our training and validation datasets were collected at different time intervals, which can simulate the datasets in prospective clinical trials and thus yield more objective results. Furthermore, the models in the two studies were only applicable to the detection of still images, which restrained the clinical application of real-time videos. Our DCNN

Table 3
Comparison between the DCNN system and endoscopists.

	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	Positive predictive value (95% CI)	Negative predictive value (95% CI)
DCNN	95.3 (93.3–96.8)	93.0 (89.3–95.5)	97.7 (95.0–99.0)	97.6 (94.8–98.9)	93.3 (89.8–95.7)
Experts	87.3 (85.2–89.3)	82.7 (75.5–89.9)	91.9 (87.2–96.6)	92.1 (88.4–95.7)	85.4 (80.1–90.7)
Trainees	73.6 (71.0–76.3)	50.2 (44.1–56.4)	97.1 (95.6–98.5)	95.1 (93.2–96.9)	66.7 (63.9–69.5)
DCNN + experts	94.3 (91.0–97.5)	97.4 (95.0–99.8)	91.1 (83.1–99.1)	92.1 (85.6–98.5)	97.9 (96.3–99.4)
DCNN + trainees	96.2 (95.8–96.7)	94.7 (93.9–95.6)	97.7 (96.8–98.6)	97.7 (96.8–98.5)	94.9 (94.1–95.7)

system is superior because it achieved excellent performance in video streams. Luo et al. developed an AI-based diagnostic system for upper gastrointestinal cancer and validated the excellent performance of this system in detecting the malignancy [16]. However, as their open data indicated, most of the images that they used contained advanced-stage GC lesions. This may explain the relatively high accuracy, sensitivity, and NPV of trainee endoscopists in their study. The training and validation datasets we used in developing our DCNN system were derived from EGC patients treated with ESD and had a histologically proven malignancy. This ensured that all cancerous images contained EGC lesions but not advanced GC lesions. Further, the sensitivity analysis showed that the DCNN system achieved comparable performance in three subgroups of EGC lesions, including intraepithelial lesions, intramucosal lesions and submucosal lesions. Moreover, because all tissues originated from ESD, the pathological diagnosis was much more convincing than that from forceps biopsy specimens. Therefore, our DCNN system is more reliable in EGC detection than previously reported systems.

To reinforce the diagnostic robustness and stability, we trained our DCNN system using reliable datasets in which labels were confirmed when more than four of five endoscopists have reached a consensus. Benefiting from the validated marks and efficient Darknet-53 model, the diagnostic robustness and stability of the DCNN system were satisfactory compared with those in previous reports. To evaluate the robustness of the DCNN system, we validated this system in one temporal validation dataset and three external validation datasets. The DCNN system achieved excellent performance in the NJDTH temporal validation dataset with an accuracy of 87.8%, a sensitivity of 95.5%, a specificity of 81.7%, and an AUC value of 0.940. In addition, the system also showed a favourable performance in the three external validation datasets. Moreover, we used Cohen kappa coefficients to assess the stability of the DCNN system and endoscopists. Our results showed that expert endoscopists achieved substantial intra-observer agreement (κ: 0.727−0.802), whereas trainee endoscopists achieved moderate intra-observer agreement (κ: 0.355−0.744). The results also revealed that the inter-observer agreement of experts was much higher than that of trainees. However, the DCNN system exhibited an extremely stable diagnostic ability, which was superior to that of endoscopists (κ: 1.000). This might be explained by the different expertise levels among the endoscopists (inter-observer agreement) and the inevitable mistakes (intra-observer agreement) due to some subjective factors. On the basis of the remarkable performance of the DCNN system, we developed an AI-based diagnostic platform to assist in the detection of EGC in routine OGD examinations. We believe that this platform could improve the detection rate of EGC and increase the accordance of diagnosis among endoscopists with different expertise levels. Multicentre prospective validation is underwent to further evaluate the assistant role of the DCNN system in EGC detection.

However, this study had several limitations. First, the DCNN system can detect only EGC and precancerous lesions under WLI but not NBI. A DCNN system that can detect lesions in NBI mode is being developed. Second, this was a retrospective study, and the excellent performance of the DCNN system cannot reflect the clinical application in the real world. We have designed a prospective randomised controlled trial to validate the applicability of this DCNN system in real-world clinical settings. Third, the DCNN system was trained and validated on images obtained using Olympus devices, which might restrain the use of other brands (e.g., Fuji) of similar devices. We will collect more images using Fuji devices in future studies. Fourth, only high-quality images were used in this project. We will collect more images with different resolutions to enhance the generalisability.

In conclusion, we have developed an efficient AI system based on DCNN for detecting EGC in real time. The DCNN system exhibited excellent performance for EGC detection in independent validation datasets and enhanced the diagnostic ability of trainee endoscopists to a level comparable to experts. However, since this study is a retrospective study, Multicentre prospective validation is needed to acquire high-level evidence for its clinical application in EGC detection.

Contributors

XZ and GX conceived and designed the study. DT, LW, QZ, YF, DZ, TL, YL, HG, XD, WZ, GX, and XZ contributed to the acquisition of data. DT, MN, GX, and XZ contributed to the analysis and interpretation of data. DT and GX drafted and reviewed the manuscript. XZ supported the project. All authors read and approved the final manuscript.

Data sharing

An open-access image database containing 300 cancerous lesions and 300 non-cancerous controls are now available on our website (http://112.74.182.39). Due to the privacy of patients, all other datasets generated or analysed in the current study are available from the corresponding author on reasonable request approved by the IRB of Nanjing University Medical School Affiliated Drum Tower Hospital (XP.Z. zouxp@nju.edu.cn).

Declaration of Competing Interests

The authors declare no competing interests.

Acknowledgements

We thank XiaMen Innovision Medical Technology for the development of the DCNN system. Employees of the company declared no competing interests; Yanxing Hu, Qi Qiu, Xianjin Yao, and Jing Feng (Innovision Medical Technology) developed the DCNN system, and Yanxing Hu provided a technical description of the development.

This project was supported by the National Natural Science Foundation of China (Grant Nos. 81672935 and 81871947), Jiangsu Clinical Medical Center of Digestive System Diseases and Gastrointestinal Cancer (Grant No. YXZXB2016002), and Nanjing Science and Technology Development Foundation (Grant No. 2017sb332019).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ebiom.2020.103146.

References

[1] 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.
[2] Katai H, Ishikawa T, Akazawa K, Isobe Y, Miyashiro I, Oda I, et al. Five-year survival analysis of surgically resected gastric cancer cases in Japan: a retrospective analysis of more than 100,000 patients from the nationwide registry of the Japanese gastric cancer association (2001-2007). Gastric Cancer 2018;21(1):144–54.
[3] Hamashima C, Systematic Review G. Guideline development group for gastric cancer screening G. Update version of the Japanese guidelines for gastric cancer screening. Jap J Clin Oncol. 2018;48(7):673–83.
[4] Ezoe Y, Muto M, Uedo N, Doyama H, Yao K, Oda I, et al. Magnifying narrowband imaging is more accurate than conventional white-light imaging in diagnosis of gastric mucosal cancer. Gastroenterology 2011;141(6):2017–25 e3.
[5] Zhang Q, Wang F, Chen ZY, Wang Z, Zhi FC, Liu SD, 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.
[6] Dohi O, Yagi N, Majima A, Horii Y, Kitaichi T, Onozawa Y, et al. Diagnostic ability of magnifying endoscopy with blue laser imaging for early gastric cancer: a prospective study. Gastric Cancer 2017;20(2):297–303.
[7] Yamada M, Oda I, Taniguchi H, Kushima R. [Chronological trend in clinicopathological characteristics of gastric cancer]. Nihon Rinsho 2012;70(10):1681–5.
[8] Zheng R, Zeng H, Zhang S, Chen W. Estimates of cancer incidence and mortality in China, 2013. Chin J Cancer 2017;36(1):66.

[9] Bi WL, Hosny A, Schabath MB, Giger ML, Birkbak NJ, Mehrtash A, et al. Artificial intelligence in cancer imaging: clinical challenges and applications. CA Cancer J Clin 2019;69(2):127–57.
[10] Milea D, Najjar RP, Zhubo J, Ting D, Vasseneix C, Xu X, et al. Artificial intelligence to detect papilledema from ocular fundus photographs. New Engl J Med 2020.
[11] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542(7639):115–8.
[12] Ting DSW, Cheung CY, Lim G, Tan GSW, Quang ND, Gan A, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA 2017;318(22):2211–23.
[13] Wu L, Zhang J, Zhou W, An P, Shen L, Liu J, et al. Randomised controlled trial of WISENSE, a real-time quality improving system for monitoring blind spots during esophagogastroduodenoscopy. Gut 2019;68(12):2161–9.
[14] Hirasawa T, Aoyama K, Tanimoto T, Ishihara S, Shichijo S, Ozawa T, et al. Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images. Gastric Cancer 2018;21(4):653–60.
[15] Yoon HJ, Kim S, Kim JH, Keum JS, Oh SI, 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).
[16] Luo H, Xu G, Li C, He L, Luo L, Wang Z, et al. Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: a multicentre, case-control, diagnostic study. Lancet Oncol 2019;20(12):1645–54.
[17] Japanese Gastric Cancer A. Japanese gastric cancer treatment guidelines 2014 (ver. 4). Gastric Cancer 2017;20(1):1–19.
[18] Van Cutsem E, Sagaert X, Topal B, Haustermans K, Prenen H. Gastric cancer. Lancet 2016;388(10060):2654–64.
[19] Yamazato T, Oyama T, Yoshida T, Baba Y, Yamanouchi K, Ishii Y, et al. Two years' intensive training in endoscopic diagnosis facilitates detection of early gastric cancer. Intern Med 2012;51(12):1461–5.
[20] Shi J, Jin N, Li Y, Wei S, Xu L. Clinical study of autofluorescence imaging combined with narrow band imaging in diagnosing early gastric cancer and precancerous lesions. J BUON 2015;20(5):1215–22.
[21] Florescu DN, Ivan ET, Ciocalteu AM, Gheonea IA, Tudorascu DR, Ciurea T, et al. Narrow band imaging endoscopy for detection of precancerous lesions of upper gastrointestinal tract. Rom J Morphol Embryol 2016;57(3):931–6.