Gastric Cancer (2022) 25:392–400
https://doi.org/10.1007/s10120-021-01256-8

ORIGINAL ARTICLE

Development and evaluation of a double-check support system using artificial intelligence in endoscopic screening for gastric cancer

Hirotaka Oura[1] · Tomoaki Matsumura[1] · Mai Fujie[2] · Tsubasa Ishikawa[1] · Ariki Nagashima[1] · Wataru Shiratori[1] · Mamoru Tokunaga[1] · Tatsuya Kaneko[1] · Yushi Imai[1] · Tsubasa Oike[1] · Yuya Yokoyama[1] · Naoki Akizue[1] · Yuki Ota[1] · Kenichiro Okimoto[1] · Makoto Arai[1] · Yuki Nakagawa[3] · Mari Inada[3] · Kazuya Yamaguchi[3] · Jun Kato[1] · Naoya Kato[1]

Received: 4 July 2021 / Accepted: 1 October 2021 / Published online: 15 October 2021
© The International Gastric Cancer Association and The Japanese Gastric Cancer Association 2021

Abstract

Background This study aimed to prevent missing gastric cancer and point out low-quality images by developing a double-check support system (DCSS) for esophagogastroduodenoscopy (EGD) still images using artificial intelligence.

Methods We extracted 12,977 still EGD images from 855 cases with cancer [821 with early gastric carcinoma (EGC) and 34 malignant lymphoma (ML)] and developed a lesion detection system using 10,994 images. The remaining images were used as a test dataset. Additional validation was performed using a new dataset containing 50 EGC and 1,200 non-GC images by comparing the interpretation of ten endoscopists (five trainees and five experts). Furthermore, we developed another system to detect low-quality images, which are not suitable for diagnosis, using 2198 images.

Results In the validation of 1983 images from the 124 cancer cases, the DCSS diagnosed cancer with a sensitivity of 89.2%, positive predictive value (PPV) of 93.3%, and an accuracy of 83.3%. EGC was detected in 93.2% and ML in 92.5% of cases. Comparing with the endoscopists, sensitivity was significantly higher in the DCSS, and the average diagnostic time was significantly shorter using the DCSS than that by the trainees. The sensitivity, specificity, PPV, and accuracy in detecting low-quality images were 65.8%, 93.1%, 79.6%, and 85.2% for "Blur" and 57.8%, 91.7%, 82.2%, and 78.1% for "Mucus adhesion," respectively.

Conclusions The DCSS showed excellent capability in detecting lesions and pointing out low-quality images.

Keywords Artificial intelligence · Gastric cancer · White light endoscopy · Screening endoscopy · Low-quality image

Introduction

According to the 2018 statistics, gastric carcinoma (GC) is the sixth leading cause of cancer death worldwide, with a high prevalence and mortality rate worldwide [1]. According to the 2017 statistics, GC is prominent in Japan with a prevalence which is second among all cancer types [2]. In a bid to reduce deaths from the disease, municipalities in Japan are required to provide mass screening for GC [3]. Barium swallowing test has been performed for decades as a screening test for GC because it has been shown to reduce GC mortality [4, 5]. Early detection of GC, including intramucosal lesions, has led to a marked reduction in mortality, while morbidity remains high [6]. Esophagogastroduodenoscopy (EGD) has been recently proven to reduce GC mortality [7–12] and has been available through Japan's screening program since 2016. The most significant problem about GC screening is missing the GC [13, 14]. The sensitivity of EGD in detecting GC has been reported to be 69–89% [15, 16]. In Japan, to prevent missing lesions, all endoscopic images of the screening test require another qualified doctor to double check the results. However, several problems have been noted with the double-checking system: it contains time-consuming and low-quality images that do not provide clear details. In general, about 40 or more EGD images are captured per case, and the number of EGD screening has

✉ Tomoaki Matsumura
matsumura@chiba-u.jp

[1] Department of Gastroenterology, Graduate School of Medicine, Chiba University, Inohana 1-8-1, Chiba 260-8670, Japan

[2] Department of Clinical Engineering Center, Chiba University Hospital, Chiba, Japan

[3] Chiba Foundation for Health Promotion and Disease Prevention, Chiba, Japan

been gradually increasing [2, 17]. This leads to a very large number of images that need to be double-checked, which is cumbersome. The detection rate of GC during mass screening is 0.63–1.28% [11, 12], which means that most of the images for double check are actually normal. The burden on doctors who perform the double check is high. In addition, some images are of low quality and do not provide clear details due to blur or mucus adhesions [18]. Such low-quality images are not suitable for screening because they interfere with the endoscopic diagnosis and lead to variability in quality between doctors and facilities. Therefore, this study aimed to develop a double-check support system (DCSS) using artificial intelligence (AI) to evaluate still images to prevent missing gastric lesions and to point out low-quality images that are not suitable for diagnosis.

Patients and methods

Study design and cases

We sought to assess the diagnostic capability of the DCSS using white light, non-magnified images by retrospectively analyzing of our database. This system uses only white light, non-magnified images. We enrolled 855 cases who underwent EGD and were diagnosed with early gastric carcinoma (EGC) or malignant lymphoma (ML) at Chiba University Hospital and Chiba Foundation for Health Promotion and Disease Prevention from September 2014 to January 2019. For GC, we included only cases in the early stage that were considered eligible for endoscopic treatment, and excluded cases in the advanced stage that were eligible for surgery or chemotherapy. Multiple endoscopists captured the endoscopic images using standard endoscopes [GIF-H260, GIF-XP260NS, GIF-H260Z, GIF-Q260J, GIF-H290, GIF-HQ290, GIF-H290Z, GIF-H290T, GIF-XP290N (Olympus Corporation, Tokyo, Japan], EG-580NW, EG-590WR, and EG-L600ZW7 (Fujifilm, Tokyo)] and standard video processors [EVIS LUCERA CV-260, CV-260SL, EVIS LUCERA ELITE CV-290 (Olympus), Advancia VP-4450, VP-4450HD, and LASEREO VP-7000 (Fujifilm)].

Annotation of lesions

We collected a series of EGD gastric images in 855 cases with EGC or ML. White light images of lesions were extracted as annotation target images. Some images used the indigo carmine dye method. We excluded both magnified and Image Enhanced Endoscopy (IEE) images. We first marked the extent of the lesions using a rectangular bounding box, then classified them into two categories ("Cancer" and "Non-cancer") to train and validate the DCSS. "Cancer" was diagnosed on pathology and was subdivided into four subclasses: "Protruding epithelial carcinoma (0-I, 0-IIa)," "Flat epithelial carcinoma (0-IIb)," "Depressed epithelial carcinoma (0-IIc)," and "Malignant lymphoma." "Non-cancer" was divided into seven subclasses: "Epithelial adenoma," "Protruding epithelial non-cancer," "Submucosal tumor," "Xanthoma," "Benign erosion," "Benign ulcer," and "Ulcer scar" (Supplementary Fig. 1; Supplementary Table 1). These subclassifications were used to aggregate the detection rates. All annotations of images were performed by three Board Certified Fellows of the Japan Gastroenterological Endoscopy Society(TM, KO, and NA).

Annotation of low-quality images

All kinds of low-quality images with or without lesions considered difficult for use in the diagnosis were also collected from the same 855 cases. Low-quality images included those of the esophagus, duodenum, and stomach. We assessed whether the images corresponded to the following three categories: "Mucus adhesion" (adhesion of mucus or residue to the surface of the stomach mucosa), "Blur" (endoscope and subject moving relatively at the moment of capture), and "Contact with lens" (adhesions to the lens surface), and annotated the corresponding items.

All annotations were performed by the three Board Certified Fellows of the Japan Gastroenterological Endoscopy Society (TM, KO, and NA).

Development of the DCSS

Based on the above-annotated data, a DCSS with the function of both lesion detection and low-quality image detection was recently developed. Cascade R-CNN [19] was used as the base model for the deep learning-based algorithm to detect lesions and Dense Net 121 [20] for that of low-quality images. CNN is one of the most widely used network models in deep learning for medical imaging. Cascade R-CNN, in particular, improves the accuracy of object detection by cascading Intersection over Union (IoU)-related thresholds [21]. Therefore, it was used in this research for the lesion detection study, which requires recognition of various gastric lesions including benign and malignant ones. On the other hand, DenseNet is a model in which each layer is connected based on the knowledge that in a convolutional network, shorter connections between the layers close to the input and the layers close to the output allow for more accurate and efficient learning [20]. DenseNet121 was adopted for the detection of the low-images in this study because it has been proven to be effective in a wide range of research purposes, not only in object detection [22–24].

Detection study

We extracted 12,977 still images from all 855 cases for the detection of lesions. From those, 10,994 images from 727 cases were randomly adopted and used as the training data set, and the rest were treated as the test data set. There were no overlap lesions between the training and test data sets. The details of the number of images are shown in Supplementary Table 1. In the test data set, the lesions were enclosed in squares and the outputs were "Cancer" or "Non-cancer."

Performance comparison study between DCSS and endoscopists

We added validation with a new data set to compare between the DCSS and ten endoscopists. A total of 50 EGC cases (25 consecutive cases with *H. pylori* present infection and 25 consecutive cases after *H. pylori* eradication) who underwent EGD in Chiba University Hospital from April 2020 were assigned, and one image containing the lesion was extracted from each case. In addition, 1200 images of the stomach were extracted from patients who underwent EGD during the same period. The 50 EGC and 1200 non-GC images were randomly sorted and verified by the DCSS, five trainees and five experts. The evaluators were informed in advance that these images included both GC and non-GC images, but were not informed of the number. All the "experts" were Board Certified Trainers of the Japan Gastroenterological Endoscopy Society with more than 10 years of experience in endoscopy, and "trainees" were those with less than four years of experience. A square was added to the area judged to be a lesion, and cases where EGC was contained inside the square and where the non-GC image was not enclosed in a square were considered correct. For the DCSS, the correct answer was given if the lesion was detected correctly, and the classifications of "Cancer" and "Non-cancer" were not taken into consideration. The time required to process each image was measured, and the average diagnostic time were calculated.

Low-quality images study

Low-quality images regardless of the presence of lesions were collected and labeled for each of the three categories. The details of the number are shown in Supplementary Table 2. For each of the "Blur", "Mucus adhesion", and "Contact with lens", a data set containing both low-quality and normal images was created, then divided into training and test data sets. There were no overlap cases between the training and test data sets. In the test data set, output if the image corresponds to each category.

Statistical analysis

In the detection study, we calculated the sensitivity, positive predictive value (PPV), and accuracy of the DCSS. In the low-quality images study, we calculated the sensitivity, specificity, PPV, negative predictive value (NPV), and accuracy of the DCSS. In the performance comparison study between DCSS and endoscopists, sensitivity, specificity, PPV, NPV, accuracy, and average diagnostic time were calculated. The McNemar test was used to compare the sensitivities, specificities, and accuracies and the Student's *t* test was used to compare average times for detection. P values < 0.05 were considered statistically significant. All statistical analyses were performed with SPSS software, version 26 (SPSS Inc., Chicago, IL, USA). The IoU (Intersection over Union) was set at 0.3.

Ethics

This study was approved by the Ethics Committee of Chiba University (approval number, 3102). All procedures were in accordance with the guidelines of the World Medical Association's Declaration of Helsinki, and informed consent was provided by all participants.

Results

Detection study

Regarding the performance of the detection of all lesions, including both "Cancer" and "Non-cancer," sensitivity, PPV, and accuracy were 86.1%, 92.5%, and 80.4%, respectively (Table 1; Fig. 1). In Table 1, "False positive" means that the DCSS detected a lesion mistakenly in the

Table 1 Confusion matrix of the double-check support system (DCSS) regarding lesion detection

DCSS diagnosis, *n*	Actual diagnosis, *n*		
	Cancer ($n=1813$)	Non-cancer ($n=259$)	False positive*
Cancer ($n=1734$)	1617	22	95
Non-cancer ($n=194$)	72	72	50
False negative**	124	165	

*"False positive" means that the DCSS detected a lesion mistakenly in the absence of a lesion

**"False negative" means that the DCSS mistakenly missed a lesion in the presence of a lesion

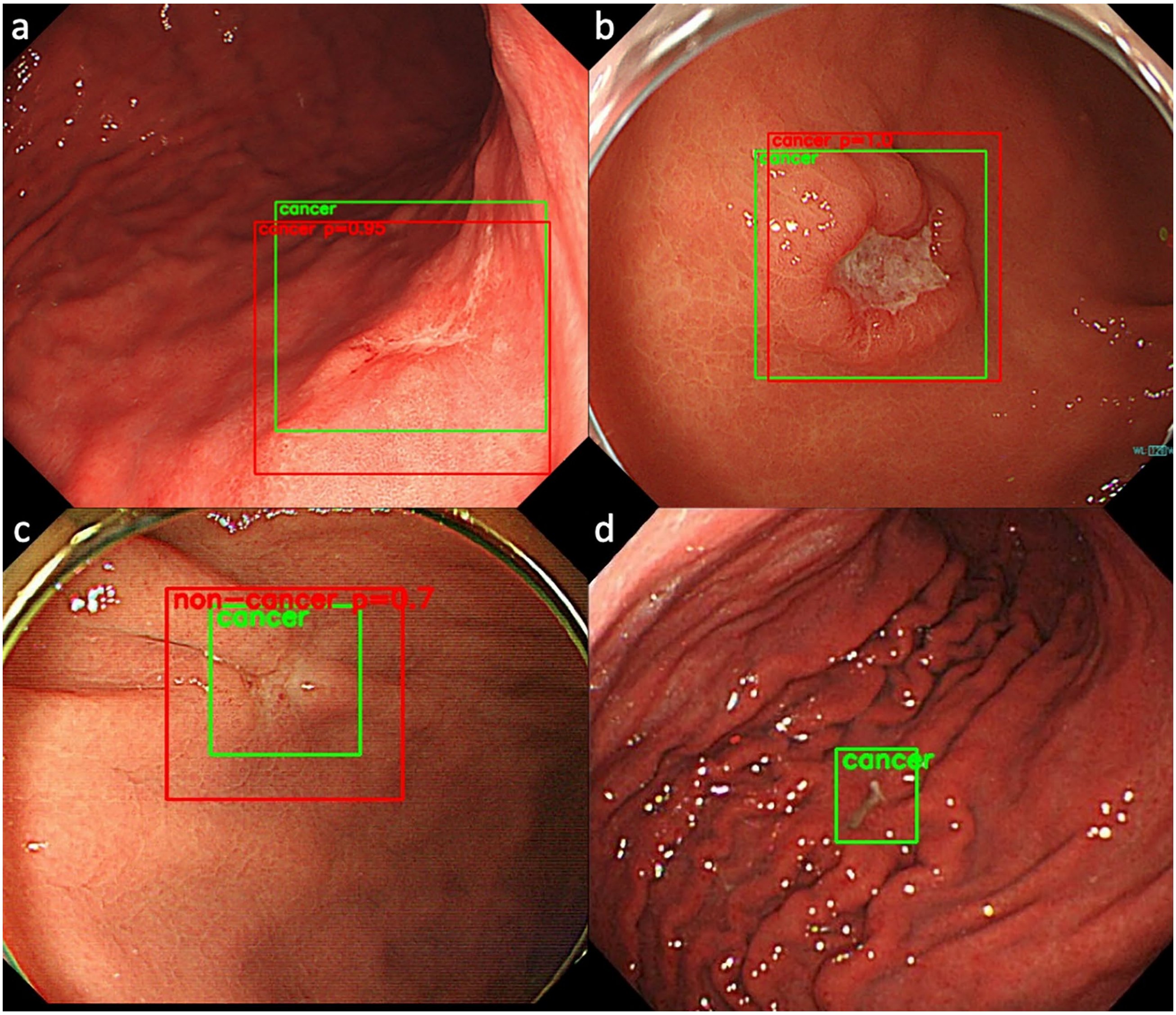

Fig. 1 Representative images of lesion detection by the double-check support system (DCSS). The green box shows the annotated area, and the red box shows the prediction by DCSS. **a** The DCSS correctly picked up and diagnosed the protruding epithelial carcinoma. **b** The DCSS correctly picked up and diagnosed the malignant lymphoma. **c** The DCSS picked up the depressed epithelial carcinoma, but misidentified it as "non-cancer." **d** The DCSS could not detect the malignant lymphoma. DCSS, Double-check support system

absence of a lesion, and "False negative" means that the DCSS mistakenly missed a lesion in the presence of a lesion. The numbers in Tables 1 and 2 are the number of bounding boxes, not the number of images. In other words, the number of bounding boxes is counted when there are multiple lesions in one image. By lesion, carcinoma was detected in 93.2% (1689/1813), adenoma in 100% (9/9), ML in 92.5% (37/40), and submucosal tumor (SMT) in 100% (2/2) (Table 2). Regarding the classification of "Cancer," the sensitivity, PPV, and accuracy were 89.2%, 93.3%, and 83.8%, respectively, while regarding the classification of "Non-cancer," they were 27.8%, 37.1%, and 18.9%, respectively.

Performance comparison study

Regarding the dataset consisting of 50 EGC and 1200 non-GC images, the accuracy, sensitivity, specificity, PPV, NPV, and average diagnostic time were 94%, 84%, 94%, 37%, 99%, and 1.86 s for the DCSS; 95%, 64%, 97%, 45%, 98%, 2.81 s for trainees, and 98%, 69%, 99%, 75%, 99%, and 2.07 s for experts. Sensitivity was significantly higher for the DCSS than for trainees and experts (84% vs. 64% and 69%). Accuracy and specificity of the DCSS were high at 94% and 94% each. However, those of endoscopists were higher than those of DCSS. The time per image was

Table 2 Sensitivity of the double-check support system regarding the detection by lesions

Annotated data, *n*			Detected images, *n*	Sensitivity, %
Cancer	Protruding epithelial carcinoma (0-I, 0-IIa)	783	736	94.0
	Flat epithelial carcinoma (0-IIb)	11	11	100
	Depressed epithelial carcinoma (0-IIc)	929	855	92.5
	Malignant lymphoma	40	37	92.5
Total	Carcinoma	1723	1602	93.0
	Cancer	1763	1639	93.0
Non-Cancer	Epithelial adenoma	9	9	100
	Epithelial non-cancer	145	59	40.7
	Submucosal tumor	2	2	100
	Xanthoma	92	45	48.9
	Benign erosion	7	5	71.4
	Benign ulcer	5	1	20.0
	Ulcer scar	45	19	42.2

significantly shorter in the DCSS than in the trainees and not significantly different in the experts (Table 3).

For the sensitivities by lesion characteristics, DCSS had significantly higher in HP-positive cases than that of experts, and had significantly higher in post-eradication cases than that of trainees and experts. By size, DCSS did not significantly differ from that of trainees and experts for lesions up to 9 mm, and was significantly higher than that found by trainees and experts for lesions between 10 and 19 mm. DCSS could detect 100% of lesions over 20 mm, while trainees could detect 81.7% and experts 73.3%. By histological type, DCSS could detect significantly more well-differentiated type lesions than either trainees or experts. For poorly differentiated type, DCSS was able to detect 100% of lesions, compared with 66.7% for trainees and 53.3% for experts. By invasion depth, DCSS could detect significantly more T1a lesions than either trainees or experts. DCSS was able to detect 100% of T1b lesions, compared with 86.7% for trainees and 66.7% for efaxperts (Table 4).

False-positive results were observed in 72 of 1,200 cases (6%) (19 cases of normal anatomical structure [cardia, pylorus, angulus], 18 cases of fold, 11 cases of gastritis [redness, atrophy, intestinal metaplasia], 7 cases of hyperplastic polyp, 4 cases of peristalsis, 3 cases of blood, and 1 case of halation, xanthoma, suction mark, foam, and extrinsic compression, each). False negatives were observed in 8 of 50 cases (16%) (5 cases of flat lesions with little tonal change from the surrounding mucosa, 2 cases of small lesions in the distant view or marginal areas of the image, and 1 case of lesion in the pyloric ring that was difficult for experts to point out (Supplementary Fig. 2).

Table 3 Diagnostic ability of the double-check support system (DCSS) compared with that of five trainees and five experts

	DCSS	Trainees	Experts	*P* value (DCSS vs. trainees)	*P* value (DCSS vs. experts)	*P* value (trainees vs. experts)
Sensitivity, % (fraction)	84.0 (42/50)	63.6 (159/250)	68.8 (172/250)	< 0.001	< 0.001	0.208
Specificity, % (fraction)	94.0 (1128/1200)	96.8 (5805/6000)	99.0 (5942/6000)	< 0.001	< 0.001	< 0.001
PPV, % (fraction)	36.8 (42/114)	44.9 (159/354)	74.8 (172/230)	–	–	–
NPV, % (fraction)	99.3 (1128/1136)	98.5 (5805/5896)	98.7 (5942/6020)	–	–	–
Accuracy, % (fraction)	93.6 (1170/1250)	95.4 (5964/6250)	97.8 (6114/6250)	< 0.001	< 0.001	< 0.001
Average time for diagnosis, s	1.86	2.81	2.07	0.006	0.333	0.052

PPV positive predictive value, *NPV* negative predictive value, *TP* true positive, *FP* false positive, *FN* false negative, *TN* true negative

Sensitivity = TP/(TP + FN); Specificity = TN/(TN + FP); PPV = TP/(TP + FP); NPV = TN/(TN + FN); Accuracy = (TP + TN)/(TP + FP + FN + TN)

Table 4 Sensitivities of the double-check support system (DCSS) compared with those of five trainees and five experts by the lesion characteristics

	DCSS	Trainees	Experts	*P* value (DCSS vs. trainees)	*P* value (DCSS vs. experts)	*P* value (trainees vs. experts)
H. pylori status						
Positive (fraction)	80.0% (20/25)	73.6% (92/125)	66.4% (83/125)	0.229	0.009	0.222
After eradication (fraction)	88.0% (22/25)	64.0% (80/125)	60.8% (76/125)	<0.001	<0.001	0.665
Size						
0–9 mm	55.5% (5/9)	60.0% (27/45)	51.1% (23/45)	0.845	0.804	0.523
10–19 mm	86.2% (25/29)	66.2% (96/145)	63.4% (92/145)	<0.001	<0.001	0.683
≥20 mm	100% (12/12)	81.7% (49/60)	73.3% (44/60)	–	–	0.302
Pathological feature						
Well differentiated type (fraction)	83.0% (39/47)	69.0% (162/235)	64.3% (151/235)	0.001	<0.001	0.278
Poorly differentiated type (fraction)	100% (3/3)	66.7% (10/15)	53.3% (8/15)	–	–	0.687
Invasion depth						
T1a (fraction)	81.8% (36/44)	66.4% (146/220)	63.2% (139/220)	<0.001	<0.001	0.510
T1b (fraction)	100% (6/6)	86.7% (26/30)	66.7% (20/30)	–	–	0.070

Low-quality image study

The sensitivity, specificity, PPV, and accuracy for the detection of low-quality images were 65.8%, 93.1%, 79.6%, and 85.2%, respectively, for "Blur"; 57.8%, 91.7%, 82.2%, and 78.1%, respectively, for "Mucus adhesion"; and 68.6%, 86.8%, 77.4%, and 79.5%, respectively, for "Contact with the lens" (Fig. 2; Table 5; Supplementary Table 3).

Discussion

Our DCSS demonstrated excellent detection capabilities for gastric lesions including malignant lesions. There are still no reports of AI algorithms that can be used especially to double check GC screening. This system could detect gastric lesions with an overall sensitivity of 86.1%, which may be helpful to prevent missing lesions in the double check of GC screening. Besides, the DCSS also showed good performance in identifying low-quality images.

DCSS showed a high sensitivity for lesion detection, 93.2% (1689/1813) of GCs were detected. During screening, a high sensitivity is required to prevent missing GC, and the DCSS is useful in this regard. In the past, Hirasawa et al. reported the world's first GC detection system using CNN [25]. A total of 13,584 GC images collected from 2,639 GCs were trained, and then validated on 2296 images including 77 GC lesions, and reported a sensitivity of 92.2% (71/77). Although the numbers were not directly comparable because we aggregated by image rather than by lesion, the results were numerically equivalent. Regarding the diagnosis of GC, our DCSS showed excellent results, with a sensitivity of 89.2%, PPV of 93.3%, and accuracy of 83.8%. In particular, the PPV of the DCSS improved markedly compared to the first report by Hirasawa et al. [25]. The reason for these excellent results can be attributed to both our use of the latest Cascade R-CNN and the fact that we trained images of non-GC lesions as "non-cancer", as Namikawa et al. reported the improvement of the overall accuracy for differentiating EGC from gastric ulcers by adding images of gastric ulcers [26]. Regarding the comparison of the capabilities of AI-based diagnostic devices and endoscopists, Ikenoyama et al. recently showed the superiority of CNN in detecting GC [27]. DCSS was also superior in detecting GC to endoscopists, including experts. The average diagnostic time was significantly shorter than that of trainees, and it is likely to be a useful tool. Regarding the diagnosis of GC, Horiuchi et al. have developed a device to differentiate EGC from gastritis in NBI magnified images, and reported high values of 95.4% sensitivity, 82.3% PPV, and 85.3% accuracy [28]. Our DCSS is superior in terms of its versatility as it uses only white light and static images, but it may be able to pick up more GCs overall when used in combination with such NBI magnification devices. In addition, in a comparison study using 50 EGC and 1200 non-GC images, the sensitivity was significantly higher than that of endoscopists. Although the PPV was not high in the comparison study, the purpose of double-checking in medical examinations is to prevent oversight, and our system is considered to be useful. For lesions other than EGC, the DCSS had a high detection sensitivity for adenomas and MLs of 100% (9/9), 92.5% (37/40), respectively. Regarding gastric adenomas, 6.8–21.4% have been reported to develop malignant transformation [29, 30]. ML is often difficult to detect in EGD [31, 32], and is also

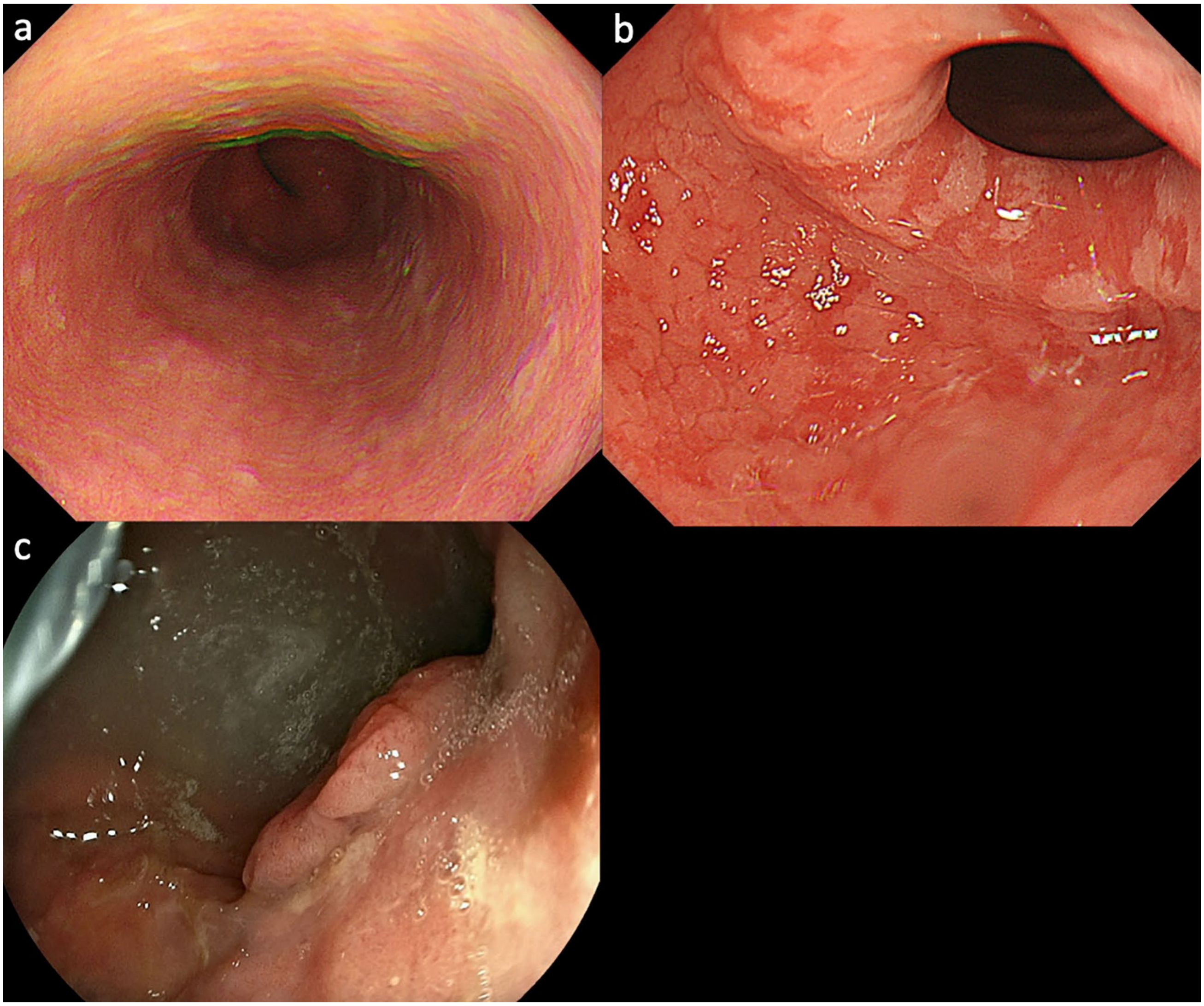

Fig. 2 Representative images of low-quality images correctly recognized by the double-check support system (DCSS). **a** The DCSS correctly identified as “Blur.” **b** the DCSS correctly identified as “Contact with lens.” **c** The DCSS correctly identified as “Mucus Adhesion”

Table 5 Statistical calculation of low-quality image detection

	Sensitivity, %	Specificity, %	PPV, %	NPV, %	Accuracy, %
Blur	65.8	93.1	79.6	86.9	85.2
Mucus	57.8	91.7	82.2	76.5	78.1
Contact with lens	68.6	86.8	77.4	80.7	79.5

PPV positive predictive value, *NPV* negative predictive value, *TP* true positive, *FP* false positive, *FN* false negative, *TN* true negative

Sensitivity = TP/(TP + FN); Specificity = TN/(TN + FP); PPV = TP/(TP + FP); NPV = TN/(TN + FN); Accuracy = (TP + TN)/(TP + FP + FN + TN)

at a risk of being missed like GC. These lesions should not be overlooked during screening for EGD. To the best of our knowledge, there are few reports that comprehensively cover lesions other than GC, and the DCSS may be useful in this regard. Another advantage is that the DCSS uses only white light still images. Since the screening is performed in a variety of facilities, spanning from high-volume centers to small local clinics, versatility is important when actually introducing the device. In this regard, the DCSS, which does not require IEE or magnification, is practical. The DCSS only uses static images; therefore, there is no need to consider when to use. The simple AI system we have developed is considered to be useful. In recent years, real-time detection devices using movies have been developed, and a sensitivity of 94.1% has been reported [33]. The values other than sensitivity are unknown, and it will take time to validate the system in actual clinical practice. On the other hand, the simple AI system we have developed is thought to be useful because it can be easily used in conjunction with the routine screening EGD that is already being performed.

The DCSS is the first device to use AI to evaluate low-quality images. The quality of images varies depending on the operators and the facilities, and there are a number of low-quality images that do not provide clear details [18]. Good quality endoscopic images that can be read in detail are essential for proper identification of lesions [34] and accurate diagnosis. However, there is no objective method to evaluate the quality of endoscopic images. Past reports using AI have excluded low-quality images by exclusive criteria and have not examined those images themselves in detail [25–28]. This study is the first attempt to use AI to evaluate low-quality images and has shown a good detection rate. Pointing out low-quality images and providing feedback to the examiner may lead to better endoscopic images.

The rate of participation in GC screening including barium swallowing and EGD in Japan has been increasing over time, rising 9.7% over the past 9 years [2]. According to previous reports [35], it takes a considerable amount of time for physicians to read endoscopic images. The diagnosis of GC in double-check screening is also expected to take a lot of time. The DCSS will reduce the burden of double-checking and prepare for future increases in the number of patients.

In the screening EGD, it is important to prevent oversight of esophageal cancer. The usefulness of AI for real-time screening of early esophageal cancer has been reported [36], and double-checking using still images may also be useful in preventing oversight. Although we did not deal with esophageal lesions in this study, it is hoped that a system which can diagnose IEE images of esophagus will be developed in the future.

This study had several limitations. First, only the cases with EGC and ML were allowed to be included in this study, and there were relatively few images of other benign lesions. The reason for this is that the DCSS is intended to improve the efficiency and accuracy of the double check in GC screening, and not primarily to distinguish between cancer and non-cancer. Therefore, the study focused on images of cancer and ML that should not be missed during screening. Secondly, advanced cancer was not included in this study. In reality, advanced cancer should be overlooked less frequently than EGC, but the overlooking of advanced cancer must be avoided, and adjustments must be made to double check the detection of advanced cancer. Third, the overall accuracy of the detection of inappropriate images is inferior. This is due to the fact that multiple physicians are annotating the images, and since the evaluation is more subjective than the annotation of lesions, variability cannot be completely eliminated. Since this is the first attempt to detect inappropriate images, it is necessary to establish and verify the uniform criteria based on a consensus among multiple physicians and facilities. Last, this study was conducted with limited cases in two institutions. It would be desirable to validate the study at a larger number of sites to reduce bias due to the medical level in the validation.

In conclusion, the DCSS has shown excellent results in detecting lesions and pointing out low-quality images, and it might help to reduce the burden on doctors in double-checking still images taken during upper gastrointestinal endoscopy.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10120-021-01256-8.

Acknowledgements This research was funded through grants (grant number, 1512) from the Chiba Foundation for Health Promotion and Disease Prevention. This research was conducted jointly with Olympus corporation. Olympus corporation provided financial support and the artificial intelligence technology to carry out the research. The authors would like to thank the staff of the Endoscopy Center of Chiba University Hospital for providing technical assistance and helping in the documentation in this study. The authors would also like to thank Enago (www.enago.jp) for English language review.

Declarations

Ethical approval This study was reviewed and approved by the Institutional Review Board of Chiba University School of Medicine.

Informed consent Informed consent was obtained from all patients to undergo the procedures involved.

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;68:394–424.
2. Cancer Registry and Statistics. Cancer Information Service, National Cancer Center, Japan (Vital Statistics of Japan). https://ganjoho.jp/reg_stat/statistics/dl/index.html#incidence4pref. Accessed 9 Mar 2021
3. Sano S, Goto R, Hamashima C. What is the most effective strategy for improving the cancer screening rate in Japan? Asian Pac J Cancer Prev. 2014;15:2607–12.
4. Oshima A, Hirata N, Ubukata T, Umeda K, Fujimoto I. Evaluation of a mass screening program for stomach cancer with a case–control study design. Int J Cancer. 1986;38:829–33.
5. Fukao A, Tsubono Y, Tsuji I, Hisamichi S, Sugahara N, Takano A. The evaluation of screening for gastric cancer in Miyagi Prefecture, Japan a population-based case–control study. Int J Cancer. 1995;60:45–8.

6. Lambert R, Guilloux A, Oshima A, Pompe-Kirn V, Bray F, Parkin M, et al. Incidence and mortality from stomach cancer in Japan, Slovenia and the USA. Int J Cancer. 2002;97:811–8.
7. Hosokawa O, Miyanaga T, Kaizaki Y, Hattori M, Dohden K, Ohta K, et al. Decreased death from gastric cancer by endoscopic screening: association with a population-based cancer registry. Scand J Gastroenterol. 2008;43:1112–5.
8. Ogura M, Hikiba Y, Maeda S, Matsumura M, Okano K, Sassa R, et al. Mortality from gastric cancer in patients followed with upper gastrointestinal endoscopy. Scand J Gastroenterol. 2008;43:574–80.
9. Hamashima C, Ogoshi K, Okamoto M, Shabana M, Kishimoto T, Fukao A. A community-based, case–control study evaluating mortality reduction from gastric cancer by endoscopic screening in Japan. PLoS ONE. 2013;8:79088.
10. Matsumoto S, Yoshida Y. Efficacy of endoscopic screening in an isolated island: a case–control study. Indian J Gastroenterol. 2014;33:46–9.
11. Hamashima C, Shabana M, Okada K, Okamoto M, Osaki Y. Mortality reduction from gastric cancer by endoscopic and radiographic screening. Cancer Sci. 2015;106:1744–9.
12. Hamashima C, Ogoshi K, Narisawa R, Kishi T, Kato T, Fujita K, et al. Impact of endoscopic screening on mortality reduction from gastric cancer. World J Gastroenterol. 2015;21:2460–6.
13. Hosokawa O, Tsuda S, Kidani E, Watanabe K, Tanigawa Y, Shirasaki S, et al. Diagnosis of gastric cancer up to three years after negative upper gastrointestinal endoscopy. Endoscopy. 1998;30:669–74.
14. Pimentaelo AR, Monteirooares M, Libânio D, Dinisibeiro M. Missing rate for gastric cancer during upper gastrointestinal endoscopy: a systematic review and meta-analysis. Eur J Gastroenterol Hepatol. 2016;28:1041–9.
15. Choi KS, Jun JK, Park EC, Park S, Jung KW, Han MA, et al. Performance of different gastric cancer screening methods in Korea: a population-based study. PLoS ONE. 2012;7:e50041.
16. Hamashima C, Okamoto M, Osaki Y, Osaki Y, Kishimoto T. Sensitivity of endoscopic screening for gastric cancer by the incidence method. Int J Cancer. 2013;133:653–60.
17. Hamashima C, Goto R. Potential capacity of endoscopic screening for gastric cancer in Japan. Cancer Sci. 2017;109:101–7.
18. Gong EJ, Lee JH, Jung K, Cho CJ, Na HK, Ahn JY, et al. Characteristics of missed simultaneous gastric lesions based on double-check analysis of the endoscopic image. Clin Endosc. 2017;50:261–9.
19. Girshick R, Donahue J, Darrell T, Malik J. Region-based convolutional networks for accurate object detection and segmentation. IEEE Trans Pattern Anal Mach Intell. 2016;38:142–58.
20. Johnson J, Karpathy A, Fei-Fei L. Densecap: Fully convolutional localization networks for dense captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. http://www.micc.unifi.it/bagdanov/pdfs/densecap.pdf. Accessed 9 Mar 2021
21. Cai Z, Vasconcelos N. Cascade RCNN: high quality object detection and instance segmentation. IEEE Trans Pattern Anal Mach Intell. 2021. https://doi.org/10.1109/TPAMI.2019.2956516.
22. Sabottke CF, Breaux MA, Spieler BM. Estimation of age in unidentified patients via chest radiography using convolutional neural network regression. Emerg Radiol. 2020;27:463–8.
23. Khan HA, Haider MA, Ansari HA, Ishaq H, Kiyani A, Sohail K, et al. Automated feature detection in dental periapical radiographs by using deep learning. Oral Surg Oral Med Oral Pathol Oral Radiol. 2021. https://doi.org/10.1016/j.oooo.2020.08.024.
24. Fujioka T, Katsuta L, Kubota K, Mori M, Kikuchi Y, Kato A, et al. Classification of breast masses on ultrasound shear wave elastography using convolutional neural networks. Ultrason Imaging. 2020;42:213–20.
25. 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;21:653–60.
26. Namikawa K, Hirasawa T, Nakano K, Ikenoyama Y, Ishioka M, Shiroma S, et al. Artificial intelligence-based diagnostic system classifying gastric cancers and ulcers: comparison between the original and newly developed systems. Endoscopy. 2020;52:1077–83.
27. Ikenoyama Y, Hirasawa T, Ishioka M, Namikawa K, Yoshimizu S, Horiuchi Y, et al. Detecting early gastric cancer: comparison between the diagnostic ability of convolutional neural networks and endoscopists. Dig Endosc. 2021;33:141–50.
28. Horiuchi Y, Hirasawa T, Ishizuka N, Tokai Y, Namikawa K, Yoshimizu S, 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;92:856–65.
29. Kolodziejczyk P, Yao T, Nakamura S, Nakamura S, Utsunomiya T, Ishikawa T, et al. Long-term follow-up study of patients with gastric adenomas with malignant transformation. An immunohistochemical and histochemical analysis. Cancer. 1994;74:2896–907.
30. Park DI, Rhee PL, Kim JE, Hyun JG, Kim YH, Son HJ, et al. Risk factors suggesting malignant transformation of gastric adenoma: univariate and multivariate analysis. Endoscopy. 2001;33:501–6.
31. Park BS, Lee SH. Endoscopic features aiding the diagnosis of gastric mucosa-associated lymphoid tissue lymphoma. Yeungnam Univ J Med. 2019;36:85–91.
32. Zullo A, Hassan C, Cristofari F, Perri F, Morini S. Gastric low-grade mucosal-associated lymphoid tissuelymphoma: Helicobacter pylori and beyond. World J Gastrointest Oncol. 2020;2:181–6.
33. Ishioka M, Hirasawa T, Tada T. Detecting gastric cancer from video images using convolutional neural networks. Dig Endosc. 2019;31:e34–5.
34. Guo R, Wang YJ, Liu M, Ge J, Zhang LY, Ma L, et al. The effect of quality of segmental bowel preparation on adenoma detection rate. BMC Gastroenterol. 2019;19:119.
35. Shichijo S, Nomura S, Aoyama K, Nishikawa Y, Miura M, Shinagawa T, et al. Application of convolutional neural networks in the diagnosis of Helicobacter pylori infection based on endoscopic images. EBioMedicine. 2017;25:106–11.
36. Guo L, Xiao X, Wu CC, Zeng X, Zhang Y, Du J, et al. Real-time automated diagnosis of precancerous lesions and early esophageal squamous cell carcinoma using a deep learning model (with videos). Gastrointest Endosc. 2020;91:41–51.

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