Digestive Endoscopy 2021; **33**: 141–150 doi: 10.1111/den.13688

Original Article

Detecting early gastric cancer: Comparison between the diagnostic ability of convolutional neural networks and endoscopists

Yohei Ikenoyama,[1,4] Toshiaki Hirasawa,[1,5] Mitsuaki Ishioka,[1] Ken Namikawa,[1] Shoichi Yoshimizu,[1] Yusuke Horiuchi,[1] Akiyoshi Ishiyama,[1] Toshiyuki Yoshio,[1,5] Tomohiro Tsuchida,[1] Yoshinori Takeuchi,[2] Satoki Shichijo,[6] Naoyuki Katayama,[4] Junko Fujisaki[1] and Tomohiro Tada[3,5]

[1]*Department of Gastroenterology, Cancer Institute Hospital, Japanese Foundation for Cancer Research,* [2]*Department of Biostatistics, School of Public Health, Graduate School of Medicine, The University of Tokyo,* [3]*AI Medical Service Inc, Tokyo,* [4]*Department of Hematology and Oncology, Mie University Graduate School of Medicine, Mie,* [5]*Tada Tomohiro Institute of Gastroenterology and Proctology, Saitama and* [6]*Department of Gastrointestinal Oncology, Osaka International Cancer Institute, Osaka, Japan*

Objectives: Detecting early gastric cancer is difficult, and it may even be overlooked by experienced endoscopists. Recently, artificial intelligence based on deep learning through convolutional neural networks (CNNs) has enabled significant advancements in the field of gastroenterology. However, it remains unclear whether a CNN can outperform endoscopists. In this study, we evaluated whether the performance of a CNN in detecting early gastric cancer is better than that of endoscopists.

Methods: The CNN was constructed using 13,584 endoscopic images from 2639 lesions of gastric cancer. Subsequently, its diagnostic ability was compared to that of 67 endoscopists using an independent test dataset (2940 images from 140 cases).

Results: The average diagnostic time for analyzing 2940 test endoscopic images by the CNN and endoscopists were 45.5 ± 1.8 s and 173.0 ± 66.0 min, respectively. The sensitivity, specificity, and positive and negative predictive values for the CNN were 58.4%, 87.3%, 26.0%, and 96.5%, respectively. These values for the 67 endoscopists were 31.9%, 97.2%, 46.2%, and 94.9%, respectively. The CNN had a significantly higher sensitivity than the endoscopists (by 26.5%; 95% confidence interval, 14.9–32.5%).

Conclusion: The CNN detected more early gastric cancer cases in a shorter time than the endoscopists. The CNN needs further training to achieve higher diagnostic accuracy. However, a diagnostic support tool for gastric cancer using a CNN will be realized in the near future.

Key words: artificial intelligence, convolutional neural network, deep learning, endoscopy, gastric cancer

INTRODUCTION

GASTRIC CANCER IS the fourth and seventh most common type of cancer in men and women, respectively, worldwide. There were over one million new cases in 2018.[1] The 5-year overall survival rates of patients in pathological stage IA was 91.5%, while it was 16.4% for patients in stage IV.[2] Therefore, endoscopic detection of gastric cancer at an early stage is important; however, it is difficult and sometimes overlooked. Several studies have reported that the false negative rate for detecting gastric cancer with esophagogastroduodenoscopy (EGD) is 4.6–25.8%.[3–11]

Recently, artificial intelligence (AI) based on deep learning through convolutional neural networks (CNNs) has made remarkable progress in various fields, including medicine. CNN is a popular deep learning method for image recognition proposed by Szegedy *et al.*[12] The use of AI in diagnosis has been previously reported.[13–22] We have shown that AI trained with endoscopic images could detect gastric cancer precisely.[23] To the best of our knowledge, studies reporting on the superiority of CNN over endoscopists in terms of diagnostic ability are limited. Therefore, we constructed a CNN using more than 13,000 images of EGD and tested it by comparing its diagnostic ability for detecting early gastric cancer with that of many endoscopists.

Corresponding: Toshiaki Hirasawa, Department of Gastroenterology Cancer Institute Hospital, 3-8-31, Ariake, Koto-ku, Tokyo 135-8550, Japan. Email: toshiaki.hirasawa@jfcr.or.jp

Received 4 December 2019; accepted 2 April 2020.

METHODS

Training dataset preparation

THE DATASET USED in this study is the same as our previous study.[23] The CNN was trained using EGD images obtained from four medical institutions (Cancer Institute Hospital Ariake, Tokyo, Japan; Tokatsu-Tsujinaka Hospital, Chiba, Japan; Tada Tomohiro Institute of Gastroenterology and Proctology, Saitama, Japan; and Lalaport Yokohama Clinic, Kanagawa, Japan) between April 2004 and December 2016. We used standard endoscopes (GIF-H290Z, GIF-H290, GIF-XP290N, GIF-H260Z, GIF-Q260J, GIF-XP260, GIF-XP260NS, and GIF-N260; Olympus Medical Systems, Tokyo, Japan) and standard endoscopic video systems (EVIS LUCERA CV-260/CLV-260 and EVIS LUCERA ELITE CV-290/CLV-290SL; Olympus Medical Systems). The EGD images were captured with standard white light imaging (WLI), chromoendoscopy using indigo carmine spraying, and narrow band imaging (NBI). Images containing poor insufflation, post-biopsy bleeding, halation, blur, defocus, or mucus were excluded from the training dataset. After selection, we used 13,584 images of 2639 gastric cancer cases as a training dataset for the CNN algorithm. These were composed of 10,474 and 3110 images of early and advanced gastric cancer, respectively (Fig. 1). These were all confirmed histologically as gastric cancer lesions using biopsy. All gastric cancer lesions in the training dataset were manually annotated with rectangular bounding boxes by an expert endoscopist (T.H.), who is also a board-certified trainer at the Japan Gastroenterological Endoscopy Society.

Training the CNN

We used the deep neural network architecture called Single Shot MultiBox Detector (SSD, https://arxiv.org/abs/1512.02325) without altering its algorithm. The AI-based diagnostic system was constructed as previously described.[23] SSD is a deep CNN that comprises 16 or more layers. To train and test the CNN, we used the Caffe deep learning framework. All layers of the CNN were fine-tuned using stochastic gradient descent with a global learning rate of 0.0001. Each image was resized to 300 × 300 pixels; the bounding box was also resized accordingly. These values were determined via trial and error to ensure all data were compatible with SSD.

Test dataset preparation

To evaluate the diagnostic accuracy of the constructed CNN, an independent test dataset comprising EGD images collected from the Cancer Institute Hospital, Tokyo, Japan, between January and May 2018 was used. We used a standard endoscope (GIF-H290Z; Olympus Medical Systems) and a standard endoscopic video system (EVIS LUCERA ELITE CV-290/CLV-290SL; Olympus Medical Systems). Only regular WLI with normal magnification were included in this study. Early gastric cancer with 20 mm maximum size was selected. We excluded postoperative stomach images, chromoendoscopy, enhanced images, such as NBI, and poor-quality images. Finally, 2940 images (209 images of 75 early gastric cancer lesions, 2731 images of non-neoplastic lesions) of 140 cases were selected; each case included 21 images (four antrum images, eight gastric

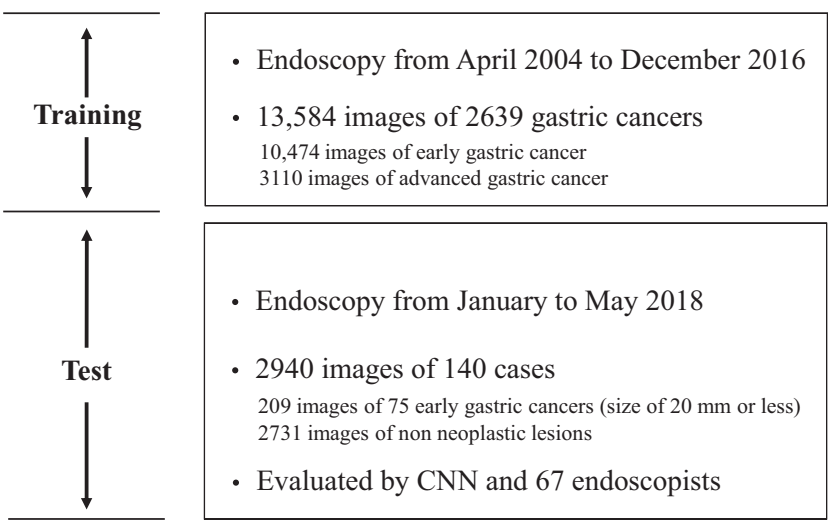

Figure 1 Patient recruitment flowchart.

body images in anterograde view, eight gastric body images in retroflexed view, and one fornix image). There were one to four images of early gastric cancer per early gastric cancer lesion (Fig. 1).

We confirmed each lesion area by comparing the endoscopic image with the resected specimen, and all gastric cancer lesions in the test dataset were manually annotated using true red rectangular bounding boxes by two experienced endoscopists (Y.I. and T.H.). The representative endoscopic images of the test dataset are shown in Figure 2. There was no overlap between the test and training datasets. Additionally, the *Helicobacter pylori* (*H. pylori*) status was confirmed using the serum anti-*H. pylori* immunoglobulin G, endoscopic signs of background mucosa atrophy, and eradication history.

Outcome measures

Per-image analysis

After training the CNN, we evaluated the diagnostic performance through the test dataset. Lesions detected by the CNN and endoscopists were indicated with green and blue rectangular frames in the endoscopic images, respectively. As the demarcation line was sometimes unclear in gastric cancer, two experienced endoscopists previously discussed dozens of other cases and defined that the CNN or endoscopists correctly detected gastric cancer lesions when the overlapped area between their rectangles and the true red rectangle was more than 40% (Fig. 3).

The CNN showed a 0–1 continuous variable number, which represented a probability score for gastric cancer in each image. Receiver operating characteristic (ROC) curves were plotted by varying the operating threshold of the probability score; the area under the curve (AUC) was then calculated. The sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of the CNN were calculated using various cut-off values for the probability score, including the score according to the Youden index, as follows.

Sensitivity. Detected number of correct gastric cancer lesions/actual number of gastric cancer lesions.

Specificity. Number of lesions that were correctly diagnosed as non-neoplastic lesions/actual number of non-neoplastic lesions.

PPV. Detected number of correct gastric cancer lesions/number of lesions diagnosed as gastric cancer by the CNN or endoscopists.

NPV. Number of lesions correctly diagnosed as non-neoplastic lesions/number of lesions diagnosed as non-neoplastic by the CNN or endoscopists.

Per-lesion analysis

When the CNN detected one gastric cancer image in multiple images of the same lesion, it was defined as a correct answer. The sensitivity of the CNN to detect gastric cancer per lesion was calculated as follows:
Sensitivity. Number of lesions correctly detected as gastric cancer lesions (one or more gastric cancer images)/actual number of gastric cancer lesions.

Comparison between the performance of CNN and endoscopists on the test dataset

To compare the diagnostic ability of the CNN and endoscopists, we recruited and divided 67 endoscopists

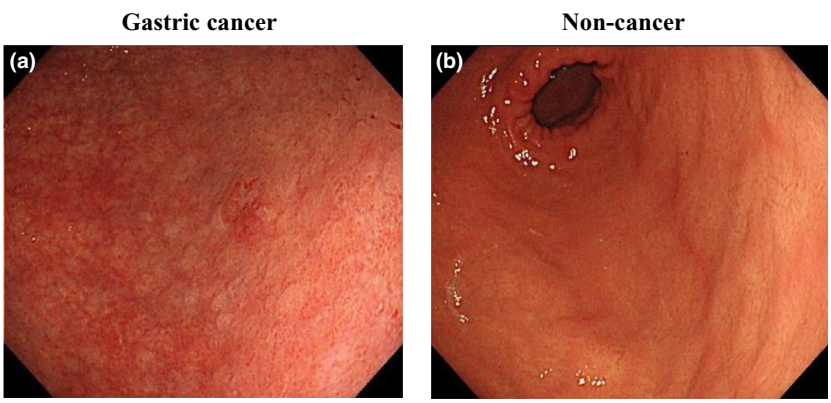

Figure 2 Representative gastric cancer and non-cancer endoscopic images. (a) A slightly reddish and depressed lesion of gastric cancer appears on the lesser curvature of the antrum. [0–IIc, 10 mm, tub1, T1a(M)]. (b) This image shows the *Helicobacter pylori* uninfected gastric mucosa. There is no cancer.

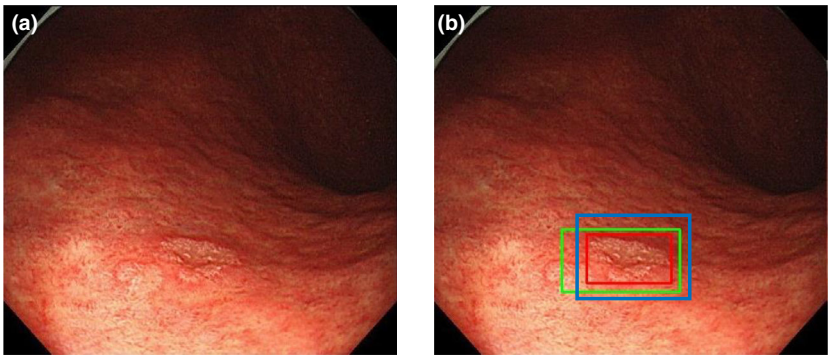

Figure 3 Definition of correct answer. (a) A reddish, depressed lesion of gastric cancer appears on the greater curvature of the lower body. [0–IIc, 9 mm, tub1, T1a(M)]. (b) The correct marking is the red rectangle. The green rectangle is the convolutional neural network (CNN) marking, and the blue rectangle is the endoscopists' marking. In this case, when the correct marking and the marking of the CNN or endoscopists overlap by 40% or more, they were judged to be correct.

into two groups: "Certified group" (comprising 33 Board certified gastroenterologists of the Japan Gastroenterological Endoscopy Society; "Non-certified group" (comprising 34 uncertified endoscopists). The mean number of endoscopy examinations for the certified and non-certified groups were 15,221 and 5465, respectively, and the mean years of experience were 18.6 and 8.2 years, respectively.

Interclass correlation coefficient (ICC) values were calculated based on a single-rating, absolute-agreement, two-way random-effects model to evaluate the interobserver variation among endoscopists. To account for the clustering of images/lesions, we estimated the sensitivity, specificity, PPV, and NPV for the CNN and endoscopists and compared these measures among groups using the generalized estimating equation method. We used R software (R Foundation for Statistical Computing, Vienna, Austria) and SAS9.4 (SAS Institute Inc., Cary, NC, USA) for statistical analyses.

Ethics

This study was approved by the Institutional Review Board of the Cancer Institute Hospital Ariake (No. 2016–1171) and Japan Medical Association (ID JMA-IIA00283).

RESULTS

Characteristics of patients and lesions in test dataset

THE PATIENT AND lesion characteristics of gastric cancer cases in the test dataset are listed in Table 1. Sixty-six lesions (88%) were mucosal cancer (T1a), and the other nine (12%) lesions were submucosal cancer. The median diameter of the tumor was 10 mm (ranging from 1.5 to 20 mm). The most macroscopic type was the superficial depressed type (0–IIc) with 64 lesions (85.3%); in terms of histopathology, 64 lesions (85.3%) were of the differentiated type. The most common *H. pylori* infection status was past infection (54 lesions, 72.0%).

Performance of CNN and endoscopists for each image

CNN

The diagnostic performance of the CNN is summarized in Table 2. The trained CNN required 45.5 ± 1.8 s to analyze the test dataset of 2940 images. Owing to the probability

Table 1 Patient and lesion characteristics of gastric cancer in test image sets

	n
Patient characteristics (*n* = 70)	
Sex, *n* (male/female)	49/21
Age, median, (range), years	68 (46–89)
Lesion characteristics (*n* = 75)	
Number of images	209
Tumor location (upper/middle/lower)	12/24/39
Tumor size, median (range), mm	10 (1.5–20)
Depth of tumor (T1a/T1b)	66/9
Macroscopic type (0-I/0-IIa/0-IIb/0-IIc/0-III)	2/8/0/65/0
Pathology (differentiated/undifferentiated)	64/11
H. pylori status (current infection/past infection/no infection)	19/54/2

H. pylori, *Helicobacter pylori*; T1a, mucosa; T1b, submucosa.

score of early gastric cancer, we evaluated the performance of the CNN per image. Figure 4 shows the ROC curves; the AUC for the CNN was 0.757, and the cut-off value for the probability score was 0.412. At the cut-off value, the sensitivity, specificity, PPV, and NPV of the CNN were 58.4% (95% confidence interval [CI], 51.4–65.1%), 87.3% (95% CI, 86.0–88.5%), 26.0% (95% CI, 22.1–30.2%), and 96.5% (95% CI, 95.8–97.2%), respectively.

Endoscopists

The diagnostic performance of the endoscopists is summarized in Table 2. The average diagnostic time was 173.0 ± 66.0 min. The overall sensitivity, specificity, PPV, and NPV were 31.9% (95% CI 28.6–35.3%), 97.2% (95% CI 96.9–97.4%), 46.2% (95% CI 41.3–51.1%), and 94.9% (95% CI 94.2–95.6%), respectively. The certified group of endoscopists had a significantly higher sensitivity (37.2% vs. 26.9%, by 10.3%; 95% CI 8.8–11.9%) and PPV (48.2% vs. 43.8%, by 4.5%; 95% CI 2.6–6.3 %) than the non-certified group. Their specificity and NPV were comparable (specificity: 97.0% vs. 97.4%, by 0.41%; 95% CI 0.3–0.6%; NPV: 95.3% vs 94.6%, by 0.7%; 95% CI 0.6–0.9%). The ICC values were 0.299 (95% CI 0.288–0.311) for all endoscopists, 0.325 (95% CI 0.312–0.338) for the certified group, and 0.284 (95% CI 0.272–0.295) for the non-certified group.

Comparison between CNN and endoscopists

The average diagnostic time for the CNN was shorter than that for the endoscopists. The sensitivity was significantly

Table 2 Diagnostic performances of CNN and endoscopists for each image

	CNN	Endoscopists		
		Certified (*n* = 33)	Non-certified (*n* = 34)	All (*n* = 67)
Diagnostic time (SD) (total)	45.5 (1.8) s	172.9 (68.4) min	173.0 (63.6) min	173.0 (66.0) min
Diagnostic time (per image)	0.0154 s	3.53 s	3.53 s	3.53 s
Sensitivity, % (95% CI)	58.4 (51.7–65.1)	37.2 (33.5–40.8)	26.9 (23.6–30.1)	31.9 (28.6–35.3)
PPV, % (95% CI)	26.0 (22.0–30.0)	48.2 (43.4–53.1)	43.8 (38.6–49.0)	46.2 (41.3–51.1)
Specificity, % (95% CI)	87.3 (86.0–88.5)	97.0 (96.7–97.2)	97.4 (97.1–97.6)	97.2 (96.9–97.4)
NPV, % (95% CI)	96.5 (95.8–97.2)	95.3 (94.6–96.0)	94.6 (93.8–95.3)	94.9 (94.2–95.6)
AUC	0.757	—	—	—

–, not applicable; AUC, area under the curve; CI, confidence interval; CNN, convolutional neural network; NPV, negative predictive value; PPV, positive predictive value; SD, standard deviation.

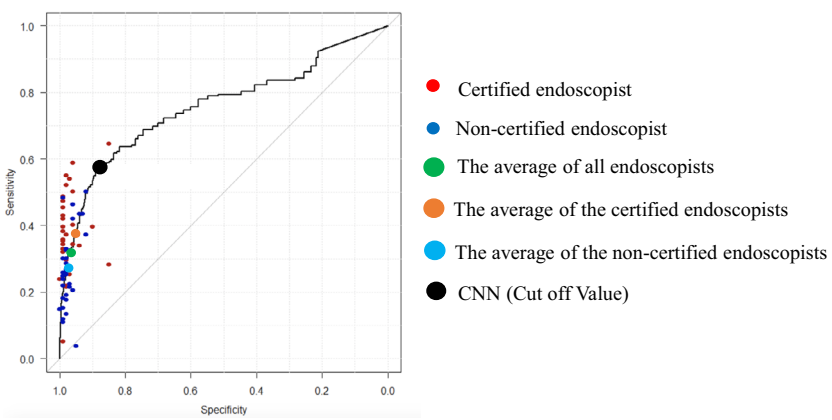

Figure 4 This graph shows the receiver operating characteristic curves for the convolutional neural network (CNN) and predictions of the endoscopists. Each endoscopist's prediction is represented by a single point. The CNN outputs a gastric cancer probability score per image, and the program then calculates a mean square of the probabilities per image. The area under the curve is 75.7%. At a cut-off value of 0.412, the sensitivity and specificity of the CNN were 58.4% and 87.3, respectively.

higher for the CNN (by 26.5%, 95% CI 14.9–32.5%), and the specificity and PPV were significantly higher for the endoscopists (specificity: by 9.9%, 95% CI 8.7–11.1%; PPV: by 20.2%, 95% CI 16.6–23.8%). Their NPV were comparable (96.5% vs. 94.9%, by 1.6%, 95% CI 1.0–2.1%). Similarly, the CNN had significantly higher sensitivity than each subgroup of endoscopists, including the certified group. Conversely, the PPV and specificity of the CNN were significantly lower than those of the non-certified group.

Performance of CNN and endoscopists for each lesion

The CNN correctly detected 60 out of 75 cases of early gastric cancers; its sensitivity was 80.0% compared with 53.4% sensitivity for the endoscopists. The sensitivity of the CNN according to tumor size, depth, macroscopic type, histopathology, and *H. pylori* infection is shown in Table 3. A total of 33 of the 36 early gastric cancer lesions (91.7%) with a diameter >10 mm were correctly detected by the CNN. In addition, the CNN could detect all lesions (9/9) with a depth of T1b.

False positives and false negatives

The causes for false positives and negatives in CNN diagnoses at a cut-off value of 0.412 are summarized in Tables 4 and 5, respectively. The most common cause of false positives was gastritis (54.8%), and the second most common cause was the misidentification of a normal anatomical structure (cardia, angulus, and pylorus). As with CNN, gastritis was the most common cause of false positives in endoscopists (73.5%), but there was no misidentification of normal anatomical structures. Representative images of false positives are shown in Figure 5.

The most common cause of false negative images was the diameter of the lesions being 10 mm or less (57.5%) (Fig. 6A). Other causes included difficult conditions, such as lesions from tangential lines (16.1%) (Fig. 6B) or lesions that were too distant (10.3%) (Fig. 6C). On the other hand, among endoscopists, gastritis was the most common cause (49.6%), and the rate of small lesions was relatively small (16.3%).

DISCUSSION

ENDOSCOPIC IMAGES OF gastric cancer vary on a case by case basis, making diagnosis difficult. Although endoscopists can undergo intensive training in EGD to drastically improve the detection rate for early gastric cancer, the training is long-term and only possible for a limited number of endoscopists.[24–26] We constructed an original CNN, trained by many gastric cancer images, that can detect more cases of early gastric cancer than experienced endoscopists. In addition, the certified group of endoscopists had a significantly higher sensitivity and PPV than the non-certified group. Therefore, the CNN may be more useful for endoscopists with limited experience, such as the non-certified group.

Table 3 Sensitivity of the CNN and endoscopists for each lesion by lesion characteristics

	Characteristics (*n*)	CNN sensitivity, % (95% CI)	Endoscopists sensitivity, % (95% CI)
Size	<10 mm (36)	91.7 (82.6–100)	62.9 (55.8–70.0)
	≦10 mm (39)	69.2 (54.8–83.7)	44.4 (35.3–53.6)
Depth	T1a (66)	77.3 (67.2–87.4)	50.6 (44.2–57.1)
	T1b (9)	100	72.8 (69.3–76.4)
Macroscopic type	0-I (2)	50 (0–100)	74.6 (43.6–100)
	0-IIa (8)	88.9 (68.4–100)	55.7 (36.1–75.4)
	0-IIc (65)	79.7 (69.8–89.5)	52.3 (45.7–58.9)
Location	Upper (12)	75.0 (50.5–99.5)	66.5 (52.5–80.6)
	Middle (24)	75.0 (57.7–92.3)	45.8 (33.8–57.8)
	Lower (39)	84.6 (73.3–99.5)	53.9 (46.1–61.6)
Histology	Differentiated type (64)	79.7 (69.8–89.5)	52.9 (46.2–59.7)
	Undifferentiated type (11)	81.8 (59.0–100)	55.5 (39.9–71.1)
H. pylori status	Current infection (19)	79.0 (60.6–97.3)	56.9 (44.7–69.1)
	Past infection (54)	79.6 (68.9–90.4)	51.8 (44.5–59.1)
	No infection (2)	100	59.7 (24.5–94.9)

CI, confidence interval; CNN, convolutional neural network; T1a, mucosa; T1b, submucosa; *H. pylori*, *Helicobacter pylori*.

Table 4 Details of false-positive images in the CNN and endoscopists diagnosis

Cause for false positives	CNN, *n* (%)	Endoscopists, *n* (%) (*n* = 67)
Total number	347	5203
Gastritis (redness, atrophy, intestinal metaplasia)	190 (54.8)	3823 (73.5)
Normal anatomical structure (cardia, pylorus, angulus)	79 (22.8)	0 (0.0)
Fold	20 (5.8)	23 (0.4)
Mucus	13 (3.7)	243 (4.7)
Halation	13 (3.7)	21 (0.4)
Scar	12 (3.5)	252 (4.8)
Foam	5 (1.4)	8 (0.2)
Blood	4 (1.2)	6 (0.1)
Vessel	2 (0.6)	138 (2.7)
Extrinsic compression	2 (0.6)	14 (0.3)
Xanthoma	2 (0.6)	103 (2.0)
Hyperplastic polyp	2 (0.6)	342 (6.6)
Submucosal tumor	1 (0.3)	178 (3.4)
Ulcer	1 (0.3)	10 (0.2)
Suction mark	1 (0.3)	42 (0.8)

CNN, convolutional neural network.

Table 5 Details of false negative images in the CNN and endoscopists diagnosis

Cause for false negatives	CNN, *n* (%)	Endoscopists, *n* (%) (*n* = 67)
Total number	87	7885
Small (≦10 mm)	50 (57.5)	1284 (16.3)
Tangential line	14 (16.1)	1089 (13.8)
Distant	9 (10.3)	1474 (18.7)
Inflammation-like	8 (9.2)	3910 (49.6)
Blood	2 (2.3)	64 (0.8)
Halation	2 (2.3)	64 (0.8)
Scar-like	2 (2.3)	0 (0.0)

CNN, convolutional neural network.

For PPV and specificity, the CNN achieved significantly lower values than those achieved by the endoscopists because the CNN was only trained for gastric cancer, therefore recognizing everything else as non-cancer. We discussed the causes of false negatives and false positives. More than half of the false negatives were lesions of 10 mm or less. However, considering that even with experienced endoscopists, the accurate diagnosis of small lesions is difficult, and that the doubling time of mucosal cancer is 2–3 years,[27] we speculate that this limitation can be clinically addressed by performing annual EGD. Other causes of false negatives included numerous tangential images and images with a distant view. There was no difference in sensitivity between the current and the past *H. pylori* infections. In other words, there was no increase in false negative cases as a result of eradication. Gastritis was a less common cause of false negatives in CNN but was the most common cause for endoscopists. This finding indicates that the CNN may be able to detect gastric cancer that endoscopists can mistake for gastritis.

The most common cause of false positives included gastritis with redness, atrophy, and intestinal metaplasia; however, it is difficult for experienced endoscopists to accurately distinguish between gastritis and gastric cancer by observing only regular WLI with normal magnification. The second cause was that the CNN misdiagnosed the anatomical structures of the cardia, pylorus, and angulus as gastric cancer. As Mori *et al*.[28] reviewed, to reduce these false positives and negatives using video-based images containing a large number of non-cancer images close to the real world seems to be effective.

In addition to a high sensitivity, the diagnostic time of the CNN was remarkably shorter than that of the endoscopists (45.5 s vs. 173.0 min). Furthermore, there was considerable interobserver variation among the endoscopists for a diagnostic agreement, which was not clearly related to expertise and experience. Regarding the ICC value of the endoscopists, we cannot determine the clear threshold to evaluate whether the reliability of the endoscopists' diagnosis was acceptable. According to a general guideline for the ICC criterion,[29] the reliability of the endoscopists' diagnosis in this study was categorized as poor. In contrast, the CNN will return consistent results as long as the thresholds for the diagnosis are not changed. This study uses verification via still images; the target of the double check system for Japanese gastric cancer screening also uses still images. Therefore, the CNN will be easily applied to this system as a supporting tool.

Our previous study has shown that CNNs can detect gastric cancer with a sensitivity of 92.2%.[23] However, the focus of our previous study was on the sensitivity of the detection of gastric cancer as a whole, including advanced gastric cancer, and it was not directly compared to that of endoscopists. The sensitivity found in this study appears to be lower than that of our previous study because of the following reasons. First, this study was limited to early gastric cancer lesions smaller than 20 mm and difficult to detect. Second, in our previous study, the sensitivity was calculated for each lesion, but not for each image; that is, if at least one gastric cancer image in multiple images of the same lesion was detected, it was counted as a correct answer. Third, in our previous study, if the lesion slightly overlapped with the marking, it was defined as correct, but

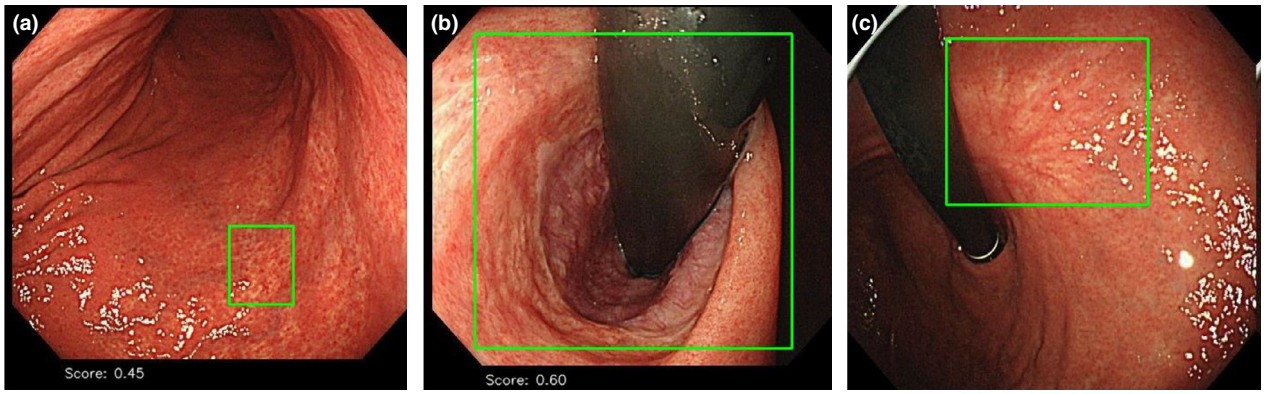

Figure 5 Representative images of false positives. The green rectangular frames show areas that the convolutional neural network misdiagnosed as gastric cancer. (a) Spotty redness associated with *Helicobacter pylori* (*H. pylori*) infection (gastritis). (b) Cardia (normal anatomical structure). (c) White scar (S2 stage) at the lesser curvature of the upper body (ulcer scar).

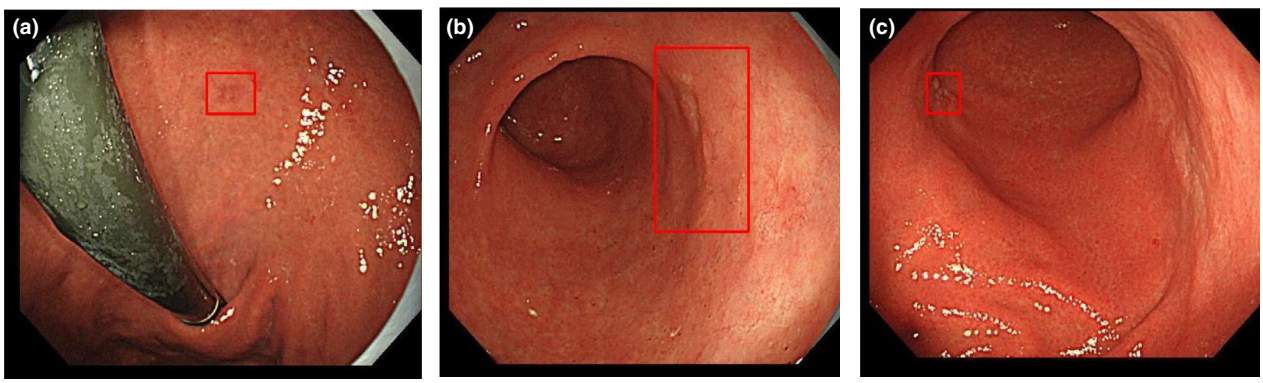

Figure 6 Representative images of false negatives. The following cancers were misdiagnosed and the assumed causes for this misdiagnosis were as follows. (a) 0– IIc, 4 mm, tub1, T1a (too small lesion). (b) Images from tangential line (tangential line). (c) Lesion at the angle captured about 7 cm away (too distant lesion).

in this study, only the one that overlapped by 40% or more was correct; lesions that were not actually recognized but accidentally marked were incorrect. However, the CNN showed a higher sensitivity than the endoscopists, including the specialists, in terms of the level of endoscopic diagnosis achieved in daily clinical practice. Furthermore, Wu *et al.*[22] compared the diagnostic abilities of endoscopists and CNN. Although the authors determined whether gastric cancer was present or absent in the image, the position and range of the cancer in the image was not evaluated in detail, and the number of endoscopists used for the comparison was small. The strength of this study is that the CNN was compared with many endoscopists, who have considerable experience, using many images under the same conditions. Therefore, we assert that the results obtained are reliable. Recently, a large multicenter study of upper gastrointestinal cancer, including gastric cancer, was published.[30] The results show a high diagnostic accuracy of over 90%, and the sensitivity was as high as that of the expert endoscopists. However, in this study, the rate of advanced gastric cancer was high and that of early gastric cancer was low (18.6%). Unlike these previous studies, our current study is limited to small early gastric cancer, which is difficult to detect. In other words, our study focuses on whether the CNN can detect cases that general endoscopists can easily overlook.

Yet, this study had several limitations. First, all test images were obtained from a single center, using the same type of endoscope (GIF-H290Z) and endoscopic video system (EVIS LUCERA ELITE CV-290/CLV-290SL). Second, we used high-quality endoscopic images for most test images. The diagnostic ability of the CNN as well as the endoscopists could be low when the conditions of the image were poor (e.g., insufficient air supply, mucus adhesion, foam, halation).[31] Third, not all false positives were histologically proven to be non-neoplastic lesions using biopsy. However, two well-experienced endoscopists

checked all test images and confirmed that there were no malignant lesions other than the target lesions. Fourth, only still images were used for both the training and test dataset in this study. Using video images may improve the performance of the CNN and represent real-life scenarios. We recently conducted a pilot study using a video image as a separate study.[19] Therein, the CNN could diagnose gastric cancer lesions with a sensitivity as high as that obtained for still images. Fifth, images obtained using chromoendoscopy or NBI were excluded from the test dataset; only those obtained using WLI were used. However, a previous study reported that image-enhanced endoscopy is rarely used unless there are suspicious findings in WLI.[32] In addition, a multicenter randomized controlled trial that examined non-magnifying NBI versus WLI revealed that there was no significant difference in the detection of gastric cancer.[33] To resolve these limitations, as Kudo *et al*.[34] also stated the importance, we are planning a multicenter prospective study using video images, including low-quality images. In addition, since the PPV and specificity of the CNN were lower than those of the endoscopists in this study, it seems that other CNNs that are updated based on false negatives and false positives may be considered.

In conclusion, we compared the diagnostic abilities of the CNN and endoscopists for detecting early gastric cancer. The CNN had a significantly higher sensitivity than experienced endoscopists, and its diagnostic time was very short. In contrast, the PPV and specificity of the CNN were lower than those of the endoscopists. This means that the diagnosis of the CNN may reduce occurrences of overlooking cancer but increase the number of biopsies for non-cancerous lesions. However, we believe that the overall diagnostic ability will be improved if endoscopists with the high PPV make a final decision on what is detected by the CNN with the high sensitivity. We expect the CNN to help detect more cases of early gastric cancer as an endoscopic support system in the near future.

ACKNOWLEDGMENTS

THE AUTHORS THANK the engineers at AI Medical Service, Inc., for their cooperation in developing the CNN.

CONFLICTS OF INTEREST

AUTHOR T.T. IS a shareholder in AI Medical Service Inc. The funding source had no role in design, practice or analysis of this study. Other authors have no COI to disclose.

FUNDING INFORMATION

NONE.

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 Katai H, Ishikawa T, Akazawa K *et al*. Five-year survival analysis of surgically resected gastric cancer cases in Japan: A retrospective analysis of more than 100,000 patients from the nationwide registry of the Japanese Gastric Cancer Association (2001–2007). *Gastric Cancer* 2018; **21**: 144–54.
3 Hosokawa O, Tsuda S, Kidani E *et al*. Diagnosis of gastric cancer up to three years after negative upper gastrointestinal endoscopy. *Endoscopy* 1998; **30**: 669–74.
4 Amin A, Gilmour H, Graham L, Paterson-Brown S, Terrace J, Crofts TJ. Gastric adenocarcinoma missed at endoscopy. *J R Coll Surg Edinb* 2002; **47**: 681–4.
5 Suvakovic Z, Bramble MG, Jones R, Wilson C, Idle N, Ryott J. Improving the detection rate of early gastric cancer requires more than open access gastroscopy: A five year study. *Gut* 1997; **41**: 308–13.
6 Yalamarthi S, Witherspoon P, McCole D, Auld CD. Missed diagnoses in patients with upper gastrointestinal cancers. *Endoscopy* 2004; **36**: 874–9.
7 Voutilainen ME, Juhola MT. Evaluation of the diagnostic accuracy of gastroscopy to detect gastric tumours: Clinico-pathological features and prognosis of patients with gastric cancer missed on endoscopy. *Eur J Gastroenterol Hepatol* 2005; **17**: 1345–9.
8 Raftopoulos SC, Segarajasingam DS, Burke V, Ee HC, Yusoff IF. A cohort study of missed and new cancers after esophagogastroduodenoscopy. *Am J Gastroenterol* 2010; **105**: 1292–7.
9 Vradelis S, Maynard N, Warren BF, Keshav S, Travis SP. Quality control in upper gastrointestinal endoscopy: Detection rates of gastric cancer in Oxford 2005–2008. *Postgrad Med J* 2011; **87**: 335–9.
10 Hosokawa O, Hattori M, Douden K, Hayashi H, Ohta K, Kaizaki Y. Difference in accuracy between gastroscopy and colonoscopy for detection of cancer. *Hepatogastroenterology* 2007; **54**: 442–4.
11 Vesey AT, Auld CD, McCole D. Missed upper gastrointestinal cancer at endoscopy: Can performance be improved by specialists? *Gut* 2012; **61**: A151–2.
12 Szegedy C, Liu W, Jia Y *et al*. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2015; 1–9.

13 Shichijo S, Nomura S, Aoyama K *et al*. Application of convolutional neural networks in the diagnosis of *Helicobacter pylori* infection based on endoscopic images. *EBioMed* 2017; **25**: 106–11.
14 Mori Y, Kudo SE, Misawa M *et al*. Real-time use of artificial intelligence in identification of diminutive polyps during colonoscopy: A prospective study. *Ann Intern Med* 2018; **169**: 357–66.
15 Takiyama H, Ozawa T, Ishihara S *et al*. Automatic anatomical classification of esophagogastroduodenoscopy images using deep convolutional neural networks. *Sci Rep* 2018; **8**: 7497.
16 Horie Y, Yoshio T, Aoyama K *et al*. Diagnostic outcomes of esophageal cancer by artificial intelligence using convolutional neural networks. *Gastrointest Endosc* 2019; **89**: 25–32.
17 Aoki T, Yamada A, Aoyama K *et al*. Automatic detection of erosions and ulcerations in wireless capsule endoscopy images based on a deep convolutional neural network. *Gastrointest Endosc* 2019; **89**: 357–63.
18 Ozawa T, Ishihara S, Fujishiro M *et al*. Novel computer-assisted diagnosis system for endoscopic disease activity in patients with ulcerative colitis. *Gastrointest Endosc* 2019; **89**: 416–21.e1.
19 Ishioka M, Hirasawa T, Tada T. Detecting gastric cancer from video images using convolutional neural networks. *Dig Endosc* 2019; **31**: e34–5.
20 Kumagai Y, Takubo K, Kawada K *et al*. Diagnosis using deep-learning artificial intelligence based on the endocytoscopic observation of the esophagus. *Esophagus* 2019; **16**: 180–7.
21 Nakagawa K, Ishihara R, Aoyama K *et al*. Classification for invasion depth of esophageal squamous cell carcinoma using a deep neural network compared with experienced endoscopists. *Gastrointest Endosc* 2019; **90**: 407–14.
22 Wu L, Zhou W, Wan X *et al*. A deep neural network improves endoscopic detection of earlygastric cancer without blind spots. *Endoscopy* 2019; **51**: 522–31.
23 Hirasawa T, Aoyama K, Tanimoto T *et al*. Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images. *Gastric Cancer* 2018; **21**: 653–60.
24 Zhang Q, Chen ZY, Chen CD *et al*. Training in early gastric cancer diagnosis improves the detection rate of early gastric cancer: An observational study in China. *Medicine (Baltimore)* 2015; **94**: e384.
25 Yamazato T, Oyama T, Yoshida T *et al*. Two years' intensive training in endoscopic diagnosis facilitates detection of early gastric cancer. *Intern Med* 2012; **51**: 1461–5.
26 Yoshida S, Yamaguchi H, Tajiri H *et al*. Diagnosis of early gastric cancer seen as less malignant endoscopically. *Jpn J Clin Oncol* 1984; **14**: 225–41.
27 Fujita S. Biology of early gastric carcinoma. *Pathol Res Pract* 1978; **163**: 297–309.
28 Mori Y, Kudo SE, Mohmed HEN *et al*. Artificial intelligence and upper gastrointestinal endoscopy: Current status and future perspective. *Dig Endosc* 2019; **31**: 378–88.
29 Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016; **15**: 155–63.
30 Luo H, Xu G, Li C *et al*. Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: A multicentre, case-control, diagnostic study. *Lancet Oncol* 2019; **20**: 1645–54.
31 Gotoda T, Uedo N, Yoshinaga S *et al*. Basic principles and practice of gastric cancer screening using high-definition white-light gastroscopy: Eyes can only see what the brain knows. *Dig Endosc* 2016; **28**(Suppl 1): 2–15.
32 Uedo N, Gotoda T, Yoshinaga S *et al*. Differences in routine esophagogastroduodenoscopy between Japanese and international facilities: A questionnaire survey. *Dig Endosc* 2016; **28** (Suppl 1): 16–24.
33 Ang TL, Pittayanon R, Lau JY *et al*. A multicenter randomized comparison between high-definition white light endoscopy and narrow band imaging for detection of gastric lesions. *Eur J Gastroenterol Hepatol* 2015; **27**: 1473–8.
34 Kudo SE, Mori Y, Misawa M *et al*. Artificial intelligence and colonoscopy: Current status and future perspectives. *Dig Endosc* 2019; **31**: 363–71.