Thieme

Diagnostic performance of artificial intelligence to identify deeply invasive colorectal cancer on non-magnified plain endoscopic images

Authors

Yuki Nakajima*,[1], Xin Zhu*,[2], Daiki Nemoto[1], Qin Li[2], Zhe Guo[2], Shinichi Katsuki[3], Yoshikazu Hayashi[4], Kenichi Utano[1], Masato Aizawa[1], Takahito Takezawa[4], Yuichi Sagara[4], Goro Shibukawa[1], Hironori Yamamoto[4], Alan Kawarai Lefor[5], Kazutomo Togashi[1]

Institutions

1 Coloproctology & Gastroenterology, Aizu Medical Center, Fukushima Medical University, Japan
2 Biomedical Information Engineering Lab, the University of Aizu, Japan
3 Gastroenterology, Otaru Ekisaikai Hospital, Japan
4 Gastroenterology, Jichi Medical University, Japan
5 Surgery, Jichi Medical University, Japan

submitted 23.4.2020
accepted after revision 24.6.2020

Bibliography
Endoscopy International Open 2020; 08: E1341–E1348
DOI 10.1055/a-1220-6596
ISSN 2364-3722

Corresponding author
Kazumoto Togashi, Aizu Medical Center, FMU – Coloproctology, 21-2 Maeda Tanisawa Kawahigashi Aizuwakamatsu , Fukushima 969-3492, Japan
Fax: 81-242752568
togashik@fmu.ac.jp

ABSTRACT

Background and study aims Colorectal cancers (CRC) with deep submucosal invasion (T1b) could be metastatic lesions. However, endoscopic images of T1b CRC resemble those of mucosal CRCs (Tis) or with superficial invasion (T1a). The aim of this study was to develop an automatic computer-aided diagnosis (CAD) system to identify T1b CRC based on plain endoscopic images.

Patients and methods In two hospitals, 1839 non-magnified plain endoscopic images from 313 CRCs (Tis 134, T1a 46, T1b 56, beyond T1b 37) with sessile morphology were extracted for training. A CAD system was trained with the data augmented by rotation, saturation, resizing and exposure adjustment. Diagnostic performance was assessed using another dataset including 44 CRCs (Tis 23, T1b 21) from a third hospital. CAD generated a probability level for T1b diagnosis for each image, and >95 % of probability level was defined as T1b. Lesions with at least one image with a probability level >0.95 were regarded as T1b. Primary outcome is specificity. Six physicians separately read the same testing dataset.

Results Specificity was 87 % (95 % confidence interval: 66–97) for CAD, 100 % (85–100) for Expert 1, 96 % (78–100) for Expert 2, 61 % (39–80) for both gastroenterology trainees, 48 % (27–69) for Novice 1 and 22 % (7–44) for Novice 2. Significant differences were observed between CAD and both novices ($P=0.013$, $P=0.0003$). Other diagnostic values of CAD were slightly lower than of the two experts.

Conclusions Specificity of CAD was superior to novices and possibly to gastroenterology trainees but slightly inferior to experts.

Introduction

Colorectal cancers (CRCs) limited to the mucosa (Tis) and also CRCs with superficial submucosal invasion (T1a) without unfavorable histology do not carry any risk of lymph node metastases [1–3]. In contrast, CRCs with deep (≥1 mm) submucosal invasion (T1b) can be associated with metastases. Recurrent lesions may develop after endoscopically removing deeply invasive CRC, especially those with unfavorable histology [1–3]. During colonoscopic examinations, therefore, it is imperative to discriminate T1b stage CRCs from less invasive ones. However, endoscopic images of T1b stage CRC resemble those of

* These authors contributed equally.

Tis/T1a stage CRC [4,5], and colonoscopists frequently have difficulty making this differentiation.

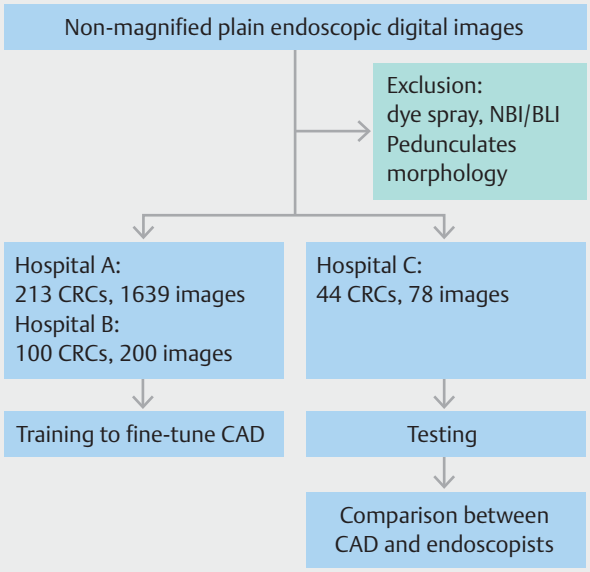
Previous reports of diagnostic performance using plain endoscopic images showed that the accuracy of identifying substantially invasive CRC was 85–91% [4,5], which is considered clinically acceptable. However, the accuracy of identifying protruding type CRCs was relatively low (82%) even for experts [5]. In spite of using narrow band imaging (NBI), recent multicenter clinical trials of non-magnified colonoscopy failed to improve diagnostic performance for T1b lesions (sensitivity 63% [6], 58% [7]). These results suggest that it is still difficult to identify T1b CRCs based on non-magnified colonoscopic observation alone. Endoscopic ultrasound or magnification colonoscopy have been used to differentiate T1b from Tis/T1a stage CRC. A miniprobe ultrasound had high accuracy (88%) in discriminating Tis from T1 CRC [8]. Pit pattern assessment using magnification colonoscopy has also resulted in a high sensitivity (85.6%) and high specificity (99.4%) [9,10]. When using endoscopic ultrasound or magnification colonoscopy, however, special devices as well as expertise is indispensable to obtain such excellent performance.

With progress in computer science, artificial intelligence (AI) is being increasingly applied to interpretation of medical images [11,12]. In colonoscopy, computer-aided diagnosis (CAD) systems for polyp detection or polyp characterization have been developed, and several clinical trials have validated the function of CAD systems [13–15]. Endocytoscopy combined with a CAD system may also play an important role in this new field [16]. Recently, Ito et al. proposed a CAD system to identify T1b CRC using plain endoscopic images, but the diagnostic performance did not have excellent results (sensitivity 89%, specificity 68%) [17]. Using narrow-band imaging (NBI), more recently, Lui T et al. reported an AI image classifier of candidates for endoscopic resection and revealed better diagnostic performance (sensitivity 88.2%. specificity 77.9%) using image-based analysis [18]. Although application of NBI for identification of T1b stage CRCs on non-magnified images may facilitate an improved performance of a CAD system, non-magnified NBI did not demonstrate excellent diagnostic performance (sensitivity 58.4%, specificity 96.4%) in the latest clinical trial [6]. Furthermore, any image-enhanced endoscopy including NBI is not so familiar to general colonoscopists and not universally used in routine colonoscopy. In this study, we aimed to develop a CAD system to differentiate T1b from Tis/T1a CRC based on non-magnified plain endoscopic images and attempted to validate its diagnostic performance.

Patients and methods

Study flow

This study was approved by the Institutional Review Board of Fukushima Medical University (registration No. 2952). To protect patient privacy, we only extracted endoscopic still images, basic endoscopic data and the final pathological diagnosis recorded in the medical information system of Aizu Medical Center Hospital (Hospital A), Jichi Medical University Hospital (Hospital B), and Otaru Ekisaikai Hospital (Hospital C). Still images from Hospitals A and B were collected until the end of September 2018 and used for training. After completing the training in mid-November 2018, still images from Hospital C were collected for the test dataset (▶ **Fig. 1**). Standards for Reporting of Diagnostic Accuracy Studies (STARD) recommendations were followed in reporting this study.



NBI: narrowband imaging, BLI: blue laser imaging, CRCs: colorectal cancers, CAD: computer-aided diagnosis

▶ **Fig. 1** Study flowchart. NBI, narrowband imaging; BLI, blue laser imaging; CRCs, colorectal cancers; CAD, computer-aided diagnosis.

Training methods

Non-magnified plain endoscopic images of early-stage CRC, deidentified and labeled only with T stage, were selected as the training dataset from existing image libraries in Hospitals A and B (▶ **Table 1**). Images enhanced with dye spray or narrow band imaging/blue laser imaging were excluded. Lesions with pedunculated morphology were also excluded because the management strategy is different from sessile lesions. Basic information regarding the lesions included depth of invasion (T stage), size and morphology. Depth of invasion follows the Japanese Classification of Colorectal, Appendiceal, and Anal Carcinoma [1] because the subclassification of T1 stage lesions is clearly defined. In this system, Tis represents cancer *in situ* (mucosal cancer), T1a represents cancer invading to submucosa less than 1 mm, T1b represents cancer invading to the submucosa more than 1 mm, and T2 represents cancer invading to the muscularis propria. Morphology was classified into polypoid (0-Is), and flat types (0-IIa, 0-IIa + IIc, 0-IIc) based on the Paris classification [19]. All endoscopic images were digitized at high resolution (1280 × 1024), using equipment from two major endoscopy manufacturers (Fujinon 72%, Olympus 28%).

In Hospital A, a total of consecutive 213 CRCs treated between May 2013 and April 2018 were extracted, and 1639 ima-

▶ **Table 1** Characteristics of lesions in the test dataset.

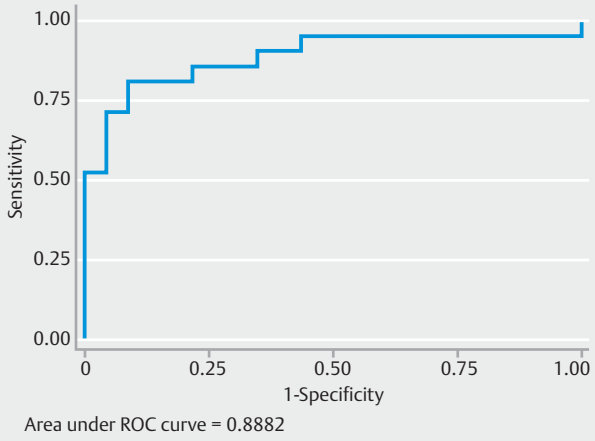
T stage, n (%)	
▪ Tis	23 (52)
▪ T1a	0 (0)
▪ T1b	21 (48)
Size, n (%)	
▪ ≤ 10 mm	3 (7)
▪ 11–20 mm	15 (34)
▪ ≥ 21 mm	26 (59)
Morphology, n (%)	
▪ Polypoid	26 (59)
▪ Flat	18 (41)

▶ **Table 2** Characteristics of lesions in the training dataset.

		Hospital A	Hospital B
T stage, n (%)	Tis	110 (52)	25 (25)
	T1a	21 (10)	25 (25)
	T1b	46 (21)	50 (50)
	T2	37 (17)	0 (0)
Size, n (%)	≤ 10 mm	58 (27)	16 (16)
	11–20 mm	75 (35)	57 (57)
	≥ 21 mm	81 (38)	27 (27)
Morphology, n (%)	Polypoid	160 (75)	50 (50)
	Flat	54 (25)	50 (50)

ges were used. Thirty-seven stage T2 CRCs less than 5 cm, labeled as T1b, were also used for training (▶ **Table 2**). In Hospital B, 100 non-consecutive CRCs, all resected by endoscopic submucosal resection between January 2016 and August 2018 were extracted, and a total of 200 images (2 images per lesion) were selected according to image quality and proportion of T staging (Tis 25 %, T1a 25 %, T1b 50 %) (▶ **Table 2**).

In this study, we used a pre-trained Resnet-50 convolutional neural network (CNN) which can output a probability level for T1b cancer. The features of CRC were calculated by a pre-trained Resnet50 [20], and then sent to a full connection and softmax layer for classification. Resnet50 has 50 layers and was pre-trained to learn rich features from over 1 million images in the ImageNet database. Resnet50 is further fine-tuned using the training dataset for the classification of deeply invasive CRC. Before training, data augmentation including rotation, saturation adjustment, resizing, and exposure adjustment were performed to increase the number of training images. Lesions were not annotated on any of the endoscopic images.



▶ **Fig. 2** Receiver operator characteristics curve for T1b lesions was calculated based on the highest confidence level of images of the lesions. The area under the curve was 0.888. The accuracy was highest (84 %) at a confidence level of 95 %.

Testing methods

Diagnostic performance was assessed using the test dataset obtained at a third hospital (Hospital C). CRCs treated between November 2017 and October 2018 were selected according to image quality and proportion of T stage (Tis/T1a 50 %, T1b 50 %).

The CAD system developed in this study generated a probability level for the diagnosis of a T1b lesion for each image. Area under the curve analysis obtained from the receiver operating characteristics (ROC) curve for T1b was calculated based on the highest probability level of images for each lesion (▶ **Fig. 2**). The area under the curve was 0.888. Considering the highest accuracy (84 %, ▶ **Supplementary Table 1**), a probability level of 95 % was selected as the optimal threshold. In lesion-based diagnosis, lesions with at least one image having a probability score > 0.95 were regarded as T1b.

Readings by endoscopists

We invited six physicians including two experts (YH, KU) in the field of colonoscopy, two gastroenterology trainees from Hospital B and two novice physicians from Hospital A to compare their diagnostic performance with the CAD system. Both expert colonoscopists have performed over 5,000 colonoscopic examinations each and are certified by Japan Gastroenterological Endoscopy Society. The two gastroenterology trainees started their training program within the last 2 years and have performed fewer than 500 colonoscopic examinations. The two novice physicians started residency training 6 months prior to this study and received 15 minutes' education by reviewing case studies just before the reading test.

On a 21-inch monitor, six physicians separately read the same test dataset, blinded to the proportion of T1b lesions. Endoscopic images identical to CAD system were presented in a random order, and physicians rated the T stage (Tis/T1a or T1b) but did not rate probability level. After completing the reading test, each image was judged as Tis/T1a or T1b, and le-

▶ **Table 3** Diagnostic performance based on lesion.

Reader	CAD	Expert 1	Expert 2	GIT 1	GIT 2	Novice 1	Novice 2
Sensitivity, n (%)	17/21 (81)	18/21 (86)	18/21 (86)	14/21 (67)	21/21 (100)[1]	21/21 (100)[1]	21/21 (100)[1]
Specificity, n (%)	20/23 (87)	23/23 (100)	22/23 (96)	14/23 (61)	14/23 (61)	11/23 (48)[1]	5/23 (22)[1]
PPV, n (%)	17/20 (85)	18/18 (100)	18/19 (95)	14/23 (61)	21/30 (70)	21/33 (64)	21/39 (54)
NPV, n (%)	20/24 (83)	23/26 (88)	22/25 (88)	14/21 (67)	14/14 (100)	11/11 (100)	5/5 (100)
Accuracy, n (%)	37/44 (84)	41/44 (93)	40/44 (91)	28/44 (64)[1]	35/44 (80)	32/44 (73)	26/44 (59)[1]

CAD, computer-aided diagnosis; GI, gastroenterology trainee; PPV, positive predictive value; NPV, negative predictive value
[1] A significant difference was observed compared with the CAD, using McNemar's test (if applicable).

sions with at least 1 image judged as T1b were regarded as T1b in lesion-based diagnosis.

Outcome measurements

The primary outcome is specificity for T1b CRCs, because specificity in the diagnosis of T1b CRC is vital for the appropriate management of patients with early stage CRC. Because radical surgery can be avoided in patients with CRCs that are not T1b lesions, it is most important that lesions that are not T1b (Tis/T1a) CRCs are identified as such. Secondary outcomes are sensitivity, positive predictive value (PPV), negative predictive value (NPV) and accuracy. Each value is expressed as frequencies of the total and percentages. The Clopper-Pearson method was used for assessing 95 % confidence intervals (CI) if necessary. McNemar's test for paired nominal data or Fisher's exact test were used to compare diagnostic values between the CAD system and physicians, as appropriate. Fisher's exact test was also used to compare proportions. Subgroup analyses by lesion size and morphology were also performed. All *P* values are two-tailed, and values <0.05 indicate statistical significance. All statistical analyses were performed with Stata 16 (Stata Corp., College Station, Texas, United State).

Results

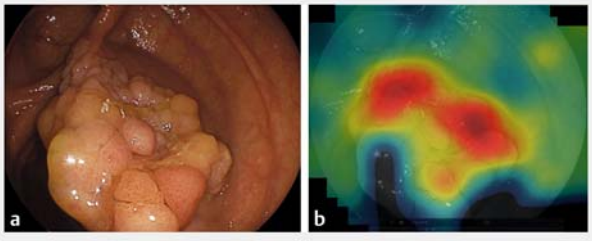
Feature of test lesions

A total of 44 CRCs (Tis 23, T1b 21) with 78 images (Tis 29, T1b 49) were used for the test. There were no significant differences in lesion size between Tis and T1b (Tis 26.0±2.0 mm, T1b 24.3±2.0 mm, $P=0.70$) lesions, but T1b lesions tended to have polypoid morphology (Tis: polypoid 43 %, flat 57 %; T1b: polypoid 76 %, flat 24 %, $P=0.036$).

Diagnostic performance of CAD and endoscopists

The specificity (primary outcome) was 87 % (95 % CI: 66–97) in CAD, 100 % (85–100) for Expert 1, 96 % (78–100) for Expert 2, 61 % (39–80) for GIT 1, 61 % (39–80) for GIT 2, 48 % (27–69) for Novice 1 and 22 % (7–44) for Novice 2 (▶ **Table 3**). Significant differences were observed between CAD and both novices ($P=0.013$, $P=0.0003$) but not in any comparisons. Sensitivity of the CAD system was slightly lower than that of the two experts (Expert 1: $P=0.414$, Expert 2: $P=0.480$). In comparison with gastroenterology trainees and novices, the sensitivity of the CAD system had a lower trend and there were significant differences with one gastroenterology trainee ($P=0.025$) and the two novice physicians (both $p=0.025$). The CAD system had accuracy inferior to both experts (Expert 1: $P=0.096$, Expert 2: $P=0.248$) but was superior to the two gastroenterology trainees and the two novices. A significantly higher accuracy of the CAD system was observed in comparison only with Novice 2 ($P=0.033$). Diagnostic performance based on image is shown in ▶ **Supplementary Table 2**. Two examples are shown in ▶ **Fig. 3** and ▶ **Fig. 4**. A class activation mapping technique which en-

▶ **Fig. 3** **a** Patient 1 had a protruding lesion with shallow depression. The convolutional neural network correctly diagnosed this lesion as Tib (confidence level 1.00), but both expert endoscopists failed to make the correct diagnosis. **b** The region of interest of the convolutional neural network was consistent with the red area of the lesion.



▶ **Fig. 4** **a** Patient 2 had a 35-mm protruding lesion. The convolutional neural network correctly diagnosed this lesion as Tis (confidence level 0.16). Both expert endoscopists were correct, but trainees and novices failed. **b** The region of interest of the convolutional neural network was consistent with the red area of the lesion.

▶ **Table 4** Diagnostic performance of CAD and experts according to lesion size and morphology.

	Reader	Size			Morphology[1]		
		≤ 20 mm	> 20 mm	*P* value[2]	Polypoid	Flat	*P* value[2]
Sensitivity	CAD	8/10 (80)	8/11 (73)	1.000	14/16 (88)	2/5 (40)	0.063
	Expert 1	10/10 (100)	8/11 (73)	0.214	14/16 (88)	4/5 (80)	1.00
	Expert 2	9/10 (90)	9/11 (82)	1.00	13/16 (81)	5/5 (100)	0.549
Specificity	CAD	6/8 (75)	14/15 (93)	0.269	2/10 (80)	12/13 (92)	0.560
	Expert 1	8/8 (100)	15/15 (100)	1.00	10/10 (100)	13/13 (100)	1.00
	Expert 2	7/8 (88)	15/15 (100)	0.348	10/10 (100)	12/13 (92)	1.00
Accuracy	CAD	14/18 (78)	22/26 (85)	0.697	22/26 (85)	14/18 (78)	0.697
	Expert 1	18/18 (100)	23/26 (88)	0.258	24/26 (92)	17/18 (94)	1.00
	Expert 2	16/18 (89)	24/26 (92)	1.00	23/26 (88)	17/18 (94)	0.634

CAD, computer-aided diagnosis
[1] Morphology is defined according the Paris classification. Polypoid includes 0-Is while flat includes 0-IIa, 0-IIa + IIc and 0-IIc
[2] Fishers' exact test

ables identification of regions of interest by CAD by displaying a red area has been developed recently. In both illustrative cases (▶ **Fig. 3** and ▶ **Fig. 4**), the regions of interest identified by CAD were consistent with the red areas on the lesions.

Subgroup analyses

Diagnostic values determined from readings by expert endoscopists based on lesion size and morphology did not show specific trends (▶ **Table 4**). The sensitivity of polypoid morphology determined by the CAD system was higher than that for flat morphology, but there was no significant difference ($P = 0.063$).

Discussion

This study demonstrates that the specificity of the CAD system trained with non-magnified plain endoscopic images was superior to gastroenterology trainees and novice physicians but slightly inferior to both experts. The sensitivity of the CAD system was nearly equivalent to both experts. The experts were likely to diagnose a lesion as a shallow CRC while the CAD system was likely to diagnose the lesion as deeply invasive CRC. A similar tendency of the CAD system was observed with gastroenterology trainees as well as novice physicians. In general, people tend to overestimate what they do not understand. The results of this study are consistent with such behavior.

In colonoscopic diagnosis of early stage CRC, high specificity for T1b is more important rather than high sensitivity although the importance of low sensitivity is not negligible. Therefore, we selected specificity as the primary outcome. One hundred percent specificity means that no unnecessary surgery was performed for patients in this series. In contrast, one hundred percent sensitivity means that all T1b CRCs were correctly identified but unnecessary surgery will be inevitable for patients with lesions incorrectly identified as T1b (false positives). A relatively lower sensitivity value will not influence the therapy for patients with deeply invasive CRC because pathological evaluation will be performed to confirm the invasiveness of the CRC after endoscopic resection. From this viewpoint, the CAD system developed in the present study achieved a better performance, compared with a previous study showing 68 % [17] and 78 % [18] specificity although these direct comparisons may be inappropriate because of using different testing datasets. Future CAD systems should aim to obtain a higher specificity comparable with the results of evaluation by experts.

To confirm the positioning of the CAD system in diagnostic performance for depth of invasion, we compared the results with readings by six endoscopists having various levels of diagnostic experience. The diagnostic performance of the CAD system was almost comparable to gastroenterology trainees. Therefore, the present CAD system cannot be adopted during colonoscopy. Furthermore, the CAD system has never been validated in real time but only using selected still images. Real-time lesion characterization may still be a distant goal. In recent practice, various modalities including endoscopic ultrasound, magnification colonoscopy and image-enhanced endoscopy have been applied to identify T1b CRCs. A future CAD system should serve to complement these existing modalities.

The current study has several advantages compared with a previous study [16, 17]. A relatively large number (1839) of endoscopic images of CRCs were collected from two hospitals for deep learning. The learning dataset included lesions treated not only by endoscopic resection but also by surgical resection. Thirty-seven stage T2 CRCs less than 5 cm were also included in the training dataset because these endoscopic images strongly resemble those of T1b CRC. This active learning [21] leads to better diagnostic performance, although the learning dataset in this study derived from data from two hospitals was completely different from the test dataset provided by a third hospital.

Diagnostic performance of the CAD system for T1b CRCs was relatively good, but the regions of interest within the image

responsible for CAD systems making the classification have been difficult to identify. Recently, the class activation mapping technique [22] has been developed, enabling one to identify the region of interest used by the CAD system by displaying a red area. Investigation of the region of interest features using class activation mapping might elucidate similarities and differences between CAD and endoscopists. As shown in ▶ **Fig. 3** and ▶ **Fig. 4**, most of the regions of interest identified by the CAD system were consistent with the areas identified by the experienced endoscopists, although the CAD system may diagnose T1b colorectal cancer using different features of the region of interest. Analysis of region of interest features using class activation mapping might lead to discovery of new endoscopic findings to indicate deep submucosal invasion.

This study has several acknowledged limitations. First, it was retrospective, and a prospective clinical trial should be performed to confirm the performance and reliability of the CAD system. Second, the datasets were not extracted based on predefined inclusion/exclusion criteria, but mainly on subjective assessment of image quality. This may lead to selection bias. For the test dataset, this also led to an imbalance in the number of images per lesion (median Tis/T1a: 1, T1b: 2). The number of images may affect diagnostic performance. Theoretically, sensitivity could increase with a larger number of images whereas specificity could decrease with a larger number of images. Third, it is not clear that the CAD system can correctly recognize T1a CRCs because the test dataset did not contain T1a lesions. This may affect its relatively good diagnostic performance. Fourth, ROC analysis was not conducted to compare between the CAD system and the endoscopists because we did not rate the confidence level of the endoscopists for individual endoscopic images. Instead, specificity was used for the comparison. A balanced analysis including sensitivity may be desirable.

Conclusion

In conclusion, the diagnostic performance of a CAD system trained with non-magnified plain endoscopic images was acceptable although inferior to readings by expert gastroenterologists. A future prospective clinical trial is warranted to confirm the performance and reliability of the CAD system.

Acknowledgements

The authors thank Dr. Yohei Funayama (gastroenterology trainee, Jichi Medical University), Dr. Takuma Kobayashi (gastroenterology trainee, Jichi Medical University), Dr. Marise Miyake (junior resident, Aizu Medical Center) and Dr. Yuki Sato (junior resident, Aizu Medical Center) for participating in the reading test.

Competing interests

The authors declare that they have no conflict of interest.

References

[1] Japanese Society for Cancer of the Colon and Rectum. Japanese Classification of Colorectal, Appendiceal, and Anal Carcinoma: the 3d English Edition [Secondary Publication]. J Anus Rectum Colon 2019; 3: 175–195

[2] Pimentel-Nunes P, Dinis-Ribeiro M, Ponchon T et al. Endoscopic submucosal dissection: European Society of Gastrointestinal Endoscopy (ESGE) guideline. Endoscopy 2015; 47: 829–854

[3] Draganov P, Wang A, Othman M et al. AGA Institute Clinical Practice Update: Endoscopic Submucosal Dissection in the United States. Clin Gastroenterol Hepatol 2019; 17: 16–25

[4] Saitoh Y, Obara T, Watari J et al. Invasion depth diagnosis of depressed type early colorectal cancers by combined use of videoendoscopy and chromoendoscopy. Gastrointest Endosc 1998; 48: 362–370

[5] Horie H, Togashi K, Kawamura YJ et al. Colonoscopic stigmata of 1 mm or deeper submucosal invasion in colorectal cancer. Dis Colon Rectum 2008: 1529–1534

[6] Backes Y, Schwartz MP, Ter Borg F et al. Multicentre prospective evaluation of real-time optical diagnosis of T1 colorectal cancer in large non-pedunculated colorectal polyps using narrow band imaging (the OPTICAL study). Gut 2019; 68: 271–279

[7] Puig I, López-Cerón M, Arnau A et al. Accuracy of the narrow-band imaging international colorectal endoscopic classification system in identification of deep invasion in colorectal polyps. Gastroenterol 2019; 156: 75–87

[8] Mukae M, Kobayashi K, Sada M et al. Diagnostic performance of EUS for evaluating the invasion depth of early colorectal cancers. Gastrointest Endosc 2015; 81: 682–690

[9] Kudo S, Tamura S, Nakajima T et al. Diagnosis of colorectal tumorous lesions by magnifying endoscopy. Gastrointest Endosc 1996; 44: 8–14

[10] Matsuda T, Fujii T, Saito TY et al. Efficacy of the invasive/non-invasive pattern by magnifying chromoendoscopy to estimate the depth of invasion of early colorectal neoplasms. Am J Gastroenterol 2008; 103: 2700–6270

[11] Kudo SE, Mori Y, Misawa M et al. Artificial intelligence and colonoscopy: Current status and future perspectives. Dig Endosc 2019; 31: 363–371

[12] Togashi K. Applications of artificial intelligence to endoscopy practice: The view from Japan Digestive Disease Week 2018. Dig Endosc 2019; 31: 270–272

[13] Misawa M, Kudo S, Mori Y et al. Artificial intelligence-assisted polyp detection for colonoscopy: initial experience. Gastroenterology 2018; 154: 2027–2029

[14] Urban G, Tripathi P, Alkayali T et al. Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy. Gastroenterology 2018; 155: 1069–1078

[15] Wang P, Berzin TM, Brown JRG et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut 2019; 68: 1813–1819

[16] Takeda K, Kudo SE, Mori Y et al. Accuracy of diagnosing invasive colorectal cancer using computer-aided endocytoscopy. Endoscopy 2017; 49: 798–802

[17] Ito N, Kawahira H, Nakashima H et al. Endoscopic diagnostic support system for cT1b colorectal cancer using deep learning. Oncology 2019; 96: 44–50

[18] Lui T, Wong K, Mak L et al. Endoscopic prediction of deeply submucosal invasive carcinoma with use of artificial intelligence. Endosc Int Open 2019; 7: E514–E520

[19] Paris endoscopic classification of superficial neoplastic lesions: esophagus, stomach, and colon: November 30 to December 1, 2002. Gastrointest Endosc 2003; 58: (Suppl. 06): S3–S43

[20] Kaiming H, Zhang X, Ren S et al. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770–778. Available from (Accessed February 12, 2020): http://openaccess.thecvf.com/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf

[21] Guo Z, Zhang R, Li Q et al. reduce false-positive rate by active learning for automatic polyp detection in colonoscopy videos. Proc IEEE International Symposium on Biomedical Imaging (ISBI'20) 2020; 01: 1655–1658

[22] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. Proc ECCV; 2014: Available from (Accessed February 12, 2020): https://cs.nyu.edu/~fergus/papers/zeilerECCV2014.pdf

▶ **Supplementary Table 1** Diagnostic performance of CAD by probability level.

Probability Level[1]	Sensitivity	Specificity	Accuracy
5%	95%	4%	45%
10%	95%	4%	45%
15%	95%	8%	48%
20%	95%	8%	48%
25%	95%	13%	50%
30%	95%	21%	55%
35%	95%	29%	59%
40%	95%	41%	66%
45%	95%	50%	70%
50%	95%	54%	73%
55%	95%	54%	73%
60%	95%	54%	73%
65%	95%	58%	75%
70%	95%	67%	80%
75%	90%	67%	80%
80%	90%	75%	82%
85%	90%	75%	82%
90%	85%	83%	84%
95%	85%	88%	84%
100%	10%	100%	59%

CAD, computer-aided diagnosis
[1] Confidence level is the threshold used in the receiver operating characteristics curve for T1b diagnosis by the computer-aided diagnosis (▶ **Fig. 2**)

▶ **Supplementary Table 2** Diagnostic performance based on image.

Reader	CAD	Expert 1	Expert 2	GIT 1	GIT 2	Novice 1	Novice 2
Sensitivity n (%)	25/49 (51)	35/49 (71)[1]	29/49 (59)	22/49 (44)	40/49 (81)[1]	46/49 (93)[1]	45/49 (91)[1]
Specificity n (%)	26/29 (90)	29/29 (100)	28/29 (96)	18/29 (62)[1]	18/29 (62)[1]	15/29 (51)[1]	8/29 (27)[1]
PPV n (%)	25/28 (89)	35/35 (100)	29/30 (96)	22/33 (66)	40/51 (78)	6/60 (76)	45/66 (68)
NPV n (%)	26/50 (52)	29/43 (67)	28/48 (58)	18/45 (40)	18/27 (66)	15/18 (83)	8/12 (66)
Accuracy n (%)	51/78 (65)	64/78 (82)[1]	57/78 (73)	40/78 (57)	58/78 (74)	61/78 (78)	53/78 (67)

CAD, computer-aided diagnosis; GIT, gastroenterology trainee; PPV, positive predictive value; NPV, negative predictive value
[1] A significant difference was observed compared with the CAD, using McNemar's test (if applicable).