Application of Convolutional Neural Networks for Detection of Superficial Nonampullary Duodenal Epithelial Tumors in Esophagogastroduodenoscopic Images

Shuntaro Inoue, MD[1], Satoki Shichijo, MD, PhD[1], Kazuharu Aoyama, MSc[2], Mitsuhiro Kono, MD[1], Hiromu Fukuda, MD[1], Yusaku Shimamoto, MD[1], Kentaro Nakagawa, MD[1], Masayasu Ohmori, MD[1], Hiroyoshi Iwagami, MD[1], Kenshi Matsuno, MD[1], Taro Iwatsubo, MD[1], Hiroko Nakahira, MD[1], Noriko Matsuura, MD[1], Akira Maekawa, MD, PhD[1], Takashi Kanesaka, MD[1], Sachiko Yamamoto, MD[1], Yoji Takeuchi, MD, PhD[1], Koji Higashino, MD[1], Noriya Uedo, MD, PhD[1], Ryu Ishihara, MD, PhD[1] and Tomohiro Tada, MD, PhD[2,3,4]

OBJECTIVES: **A superficial nonampullary duodenal epithelial tumor (SNADET) is defined as a mucosal or submucosal sporadic tumor of the duodenum that does not arise from the papilla of Vater. SNADETs rarely metastasize to the lymph nodes, and most can be treated endoscopically. However, SNADETs are sometimes missed during esophagogastroduodenoscopic examination. In this study, we constructed a convolutional neural network (CNN) and evaluated its ability to detect SNADETs.**

METHODS: **A deep CNN was pretrained and fine-tuned using a training data set of the endoscopic images of SNADETs (duodenal adenomas [N = 65] and high-grade dysplasias [HGDs] [N = 31] [total 531 images]). The CNN evaluated a separate set of images from 26 adenomas, 8 HGDs, and 681 normal tissue (total 1,080 images). The gold standard for both the training data set and test data set was a "true diagnosis" made by board-certified endoscopists and pathologists. A detected tumor was marked with a rectangular frame on the endoscopic image. If it overlapped at least a part of the "true tumor" diagnosed by board-certified endoscopists, the CNN was considered to have "detected" the SNADET.**

RESULTS: **The trained CNN detected 94.7% (378 of 399) of SNADETs on an image basis (94% [280 of 298] of adenomas and 100% [101 of 101] of HGDs) and 100% on a tumor basis. The time needed for screening the 399 images containing SNADETs and all 1,080 images (including normal images) was 12 and 31 seconds, respectively.**

DISCUSSION: **We used a novel algorithm to construct a CNN for detecting SNADETs in a short time.**

INTRODUCTION

Small bowel adenocarcinomas are generally defined as duodenal adenocarcinomas excluding ampullary carcinoma, jejunal adenocarcinoma, and ileal adenocarcinoma. They account for $<0.5\%$ of all malignant tumors and $<5\%$ of all malignant tumors of the gastrointestinal tract (1). The annual incidence of small bowel adenocarcinomas in the West is extremely low at 2.2–5.7 per million individuals (1); thus, it is considered a rare cancer. Duodenal adenocarcinoma accounts for 45% of small bowel adenocarcinomas, and its 5-year survival rate is the lowest among all malignant small bowel tumors ($<30\%$) (2,3). If diagnosed at an advanced stage, highly invasive treatments, such

as pancreaticoduodenectomy, become necessary. When advanced carcinoma is unresectable, the prognosis is poor.

A superficial nonampullary duodenal epithelial tumor (SNADET) is defined as a mucosal or submucosal sporadic tumor of the duodenum that does not arise from the papilla of Vater. Because SNADETs rarely metastasize, most can be cured by less invasive treatment, such as endoscopic resection. The detection of SNADETs is reportedly increasing with the recent widespread use of endoscopy. However, this disease can be easily missed because it is usually flat and exhibits minimal surface change. This fact is supported by the large discrepancy in the detection rates of SNADETs among previously published

reports (0.1%–3.0%) (4–8). A potential solution to mitigate the difficulty in the endoscopic detection of SNADETs is the application of computer-aided diagnosis.

The role of artificial intelligence (AI) using deep learning in various medical fields, particularly radiation oncology (9), skin cancer classification (10), and diabetic retinopathy (11), has been described in the literature. More recently, the use of convolutional neural networks (CNNs) has made it possible to diagnose esophageal cancer (12–14), diagnose *Helicobacter pylori* gastritis (15,16), detect gastric cancer (17), classify anatomical locations in the esophagogastroduodenoscopic images (18), and evaluate the activity of ulcerative colitis (19). Deep learning is a powerful machine learning technique that can interpret medical images based on a series of proprietary algorithms developed by accumulated data. Deep learning can learn expressions of data at multiple levels of abstraction using a computational model consisting of multiple processing layers (20). In this study, we constructed a diagnostic system using the endoscopic images and evaluated whether the CNN could identify SNADETs based on the endoscopic images.

METHODS

Preparation of data set
Esophagogastroduodenoscopy was performed as a screening or pretreatment examination in clinical practice, and the images were collected using endoscopes (GIF-H290Z, GIF-H290, GIF-H260Z, and GIF-Q240Z; Olympus Medical Systems, Tokyo, Japan). We prepared data (development data set) for education and construction of the AI-based diagnostic system. We retrospectively reviewed the esophagogastroduodenoscopic images obtained from 108 patients (130 tumors) from August 2016 to November 2018 at the Osaka International Cancer Institute. The endoscopic images in the former period (from August 2016 to June 2018) were used as the training data set, and the images in the latter period (from July to November 2018) were used as the test data set. We collected 1,546 training images of 96 SNADETs diagnosed histologically as either high-grade dysplasia (HGD) (N = 31) or adenoma (N = 65). The pathological diagnosis was made by board-certified pathologists or under the supervision of board-certified pathologists. Most of the diagnoses were based on resected specimens, whereas some were based on biopsy specimens. All the images of SNADETs were manually marked by a board-certified endoscopist (S.I.) and confirmed by another board-certified endoscopist (S.S.). The endoscopic images of SNADETs were further screened to exclude obscure images caused by halation, blurring, defocus, mucus, food residue, and bleeding after biopsy. Finally, 531 images containing SNADETs (249 white-light images, 90 indigo carmine images, and 192 narrow-band images) were prepared for the development data set. No normal images without a tumor were included in the development data set. The number of images used in the development data set was relatively smaller than previous studies (13–19) because of the rarity and uniformity of the SNADETs.

Another data set (test data set) was prepared to evaluate the diagnostic accuracy of the constructed CNN. The test data set comprised 399 images obtained from 34 tumors (8 HGDs and 26 adenomas; 141 white-light imaging [WLI], 61 indigo carmine imaging, and 197 narrow-band imaging [NBI]) and 681 normal images (573 WLI and 108 NBI). No overlapping exists between the development data set and test data set.

Training algorithm
The mathematical algorithm used was a CNN called the Single-Shot Multibox Detector (https://arxiv.org/abs/1512.02325) comprising ≥16 layers. We used the Single-Shot Multibox Detector without altering its algorithm. When a CNN is trained with high massive numbers of high-quality images, the CNN model will acquire excellent performance to detect diseases or objects. All the regions showing SNADETs in the training image set were manually annotated by an experienced endoscopist with rectangular bounding boxes. The annotation was executed separately, and each image was double-checked.

The CNN was trained by the data set through the Caffe deep learning framework. This framework, which was used to train and validate the CNN in this study, is one of the most popular and widely used frameworks and was originally developed at the Berkeley Vision and Learning Center. The CNN was trained to recognize the areas within the bounding boxes as a representative of SNADETs and the other areas as a representative of the background. Each image was resized to 300 × 300 pixels to match the input size of the CNN, and the bounding box was also resized accordingly. All layers of the CNN were fine-tuned from weights of ImageNet using stochastic gradient descent as a back-propagation method with a global learning rate of 0.0001 and a batch size of 32. To acquire a high-performance CNN model, we must find proper values for hyperparameters, such as the learning rate and weight decay. The values were established by repeated trial and error.

Outcome measures of AI diagnosis
After the CNN was constructed using the training image set, the performance of the CNN was evaluated using the independent test images prepared as the validation data set. When the trained CNN detected a SNADET from the input data of the test image, a diagnosis (HGD or adenoma) was made, and a rectangular frame was displayed within the endoscopic image to surround the lesion of interest. The cutoff value of the probability score was set at 0.4; a score of <0.4 was judged as negative even if it detected the lesion.

If the CNN recognized even a part of a SNADET, it was considered able to accurately detect the lesion. When the CNN did not recognize a SNADET in an image containing a SNADET, the result was judged as false-negative. When the CNN diagnosed nontumor structures as SNADETs, the result was judged as false-positive. The primary outcome measures were accuracy, sensitivity, specificity, the positive predictive value (PPV), and the negative predictive value (NPV). The sensitivity of the CNN was calculated by dividing the number of lesions diagnosed exactly as SNADETs by the actual number of SNADETs. The specificity was calculated by dividing the number of images that AI correctly diagnosed as non-SNADETs by the total number of non-SNADET images. The PPV was calculated by dividing the number of images that the CNN correctly diagnosed as SNADETs by all the images diagnosed by the CNN as SNADETs. Finally, the NPV was calculated by dividing the number of images that the CNN accurately diagnosed as non-SNADETs by the number of all the images diagnosed by the CNN as non-SNADETs. We used R software version 3.5.1 for all statistical analyses, and a P value of <0.05 was considered statistically significant.

Table 1. Characteristics of training image set and test image set

	Training image set	Test image set
Sex		
Male	57	22
Female	30	9
Median age, yr (range)	66 (35–88)	70 (48–90)
Median tumor size, mm (range)	10 (2–50)	12 (3–50)
Tumor location		
Bulbs	17	4
Second portion	79	30
Macroscopic type		
0–I	14	3
0–IIa	62	25
0–IIc	14	4
0–IIa + IIc	6	2
Histopathology		
Adenoma	65	26
High-grade dysplasia	31	8

Ethical approval

This study was approved by the Institutional Review Board of the Osaka International Cancer Institute (no. 2017–1710059178) and the Japan Medical Association (ID: JMA-IIA00283).

RESULTS

Table 1 shows the characteristics of the 118 patients and 130 lesions used for the training image set and test image set. In the test image set, 8 lesions (25%) were HGDs and 26 (75%) were adenomas. The median tumor diameter was 12 mm (range, 3–50 mm). The trained CNN diagnosed a total of 1,080 images acquired from 34 SNADETs (399 images) and the normal duodenum (681 images). The trained CNN detected 94.7% (378 of 399 images) of SNADETs on an image basis (94% [280 of 298] of adenomas and 100% [101 of 101] of HGDs) and 100% on a lesion basis. Despite the inclusion of 5 lesions of ≤5 mm, all lesions were detected by the CNN. Figure 1 shows a small lesion of 3 mm that was detected by the CNN not only in a close-up image but also in an image relatively distant from the lesion. The time taken by the CNN to diagnose the 399 images of SNADETs and all 1,080 images (including normal images) was 12 and 31 seconds, respectively. The detailed results of the AI diagnosis are shown in Table 2. The sensitivity and specificity of the AI diagnosis was 94.7% (378 of 399) and 87.4% (596 of 681), respectively. The PPV and NPV was 80.8% and 97.4%, respectively.

The rate of false-positive results was 12.6% (86 of 681 normal images). False-positives were caused by normal duodenal folds in 45 images, normal duodenal mucosa in 23 images, duodenal papillary folds in 9 images, and low-quality images (e.g., halation) in 9 images (Figure 2a–d). The rate of false-negative results was 5.3% (21 of 399 images). Most of these false-negatives were caused by lesion imaging at a distance (Figure 3).

A comparison of the diagnostic results between WLI and NBI is shown in Table 3. The sensitivity of NBI was significantly higher than that of WLI ($P = 0.009$). By contrast, the specificity was significantly higher with WLI than that with NBI ($P = 0.001$) (Figure 4). The CNN detected all the HGDs but not all adenomas. Higher detection rate of HGDs was because of its large sizes (12–20 mm) compared with adenomas (12 of 34 were less than 10 mm).

DISCUSSION

In this study, we built an AI-based diagnostic imaging system that detects SNADETs using a CNN that was trained using the images of SNADETs. The results demonstrated that the CNN was able to detect SNADETs with high accuracy.

To the best of our knowledge, this is the first study to verify the performance of an AI system in relation to SNADETs. The overall SNADET detection sensitivity was 100% on a lesion basis and 94.7% on an image basis. Our AI system successfully detected 5 small SNADETs of ≤5 mm even from a distance, suggesting the high detectability of our system. However, some normal duodenal structures were diagnosed as SNADETs at a false-positive rate of 12.6%. Most false-positives were caused by a misinterpretation of a peristalsis-associated raised fold as a lesion. Whether such errors can be resolved by training the system using a higher number of normal images requires further study.
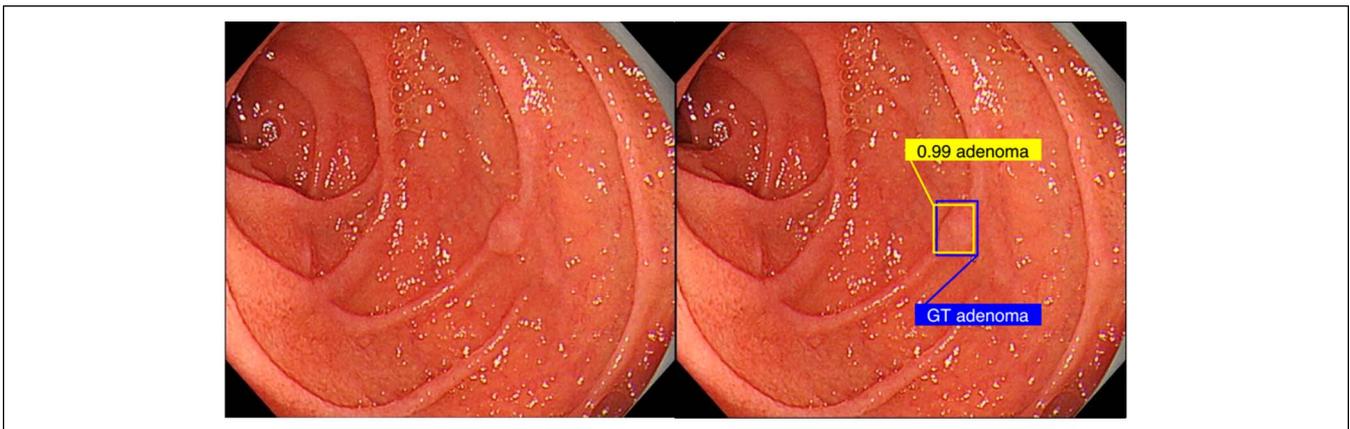

Figure 1. Small lesion of 3 mm in diameter. The CNN was able to detect this small lesion not only in near images but also in images that were relatively far away. CNN, convolutional neural network; GT, ground truth.

ENDOSCOPY

Table 2. Detailed results of AI diagnosis

Sensitivity	94.7% (378/399)
Specificity	87.4% (595/681)
PPV	80.8% (383/474)
NPV	97.4% (595/611)

AI, artificial intelligence; NPV, negative predictive value; PPV, positive predictive value.

In this study, NBI had significantly higher sensitivity and lower specificity for SNADETs than did WLI. Considering that NBI enhances the surface structure, it may increase the sensitivity for SNADETs. However, these results may be biased by factors, such as the image condition and training data set. Screening of the duodenum is usually performed by WLI, which is sometimes followed by a detailed observation by NBI. Accordingly, most WLI images are taken from a distance, whereas most NBI images are taken from a nearby location. Such a difference in imaging conditions may have caused the difference in sensitivity. In addition, WLI usually includes the surrounding mucosa because the images are taken from a distance, whereas NBI mainly includes the diseased area. The lack of normal structures in the training data set of NBI may have caused the low specificity. Considering these biases, further investigation is required to determine the relative usefulness of NBI and WLI for the detection of SNADETs.

This study has several limitations. First, all the images used to train and verify the performance of the CNN came from a single facility. SNADETs are rare compared with gastric, colorectal, and esophageal tumors; thus, only a limited number of training images could be obtained from a single facility.
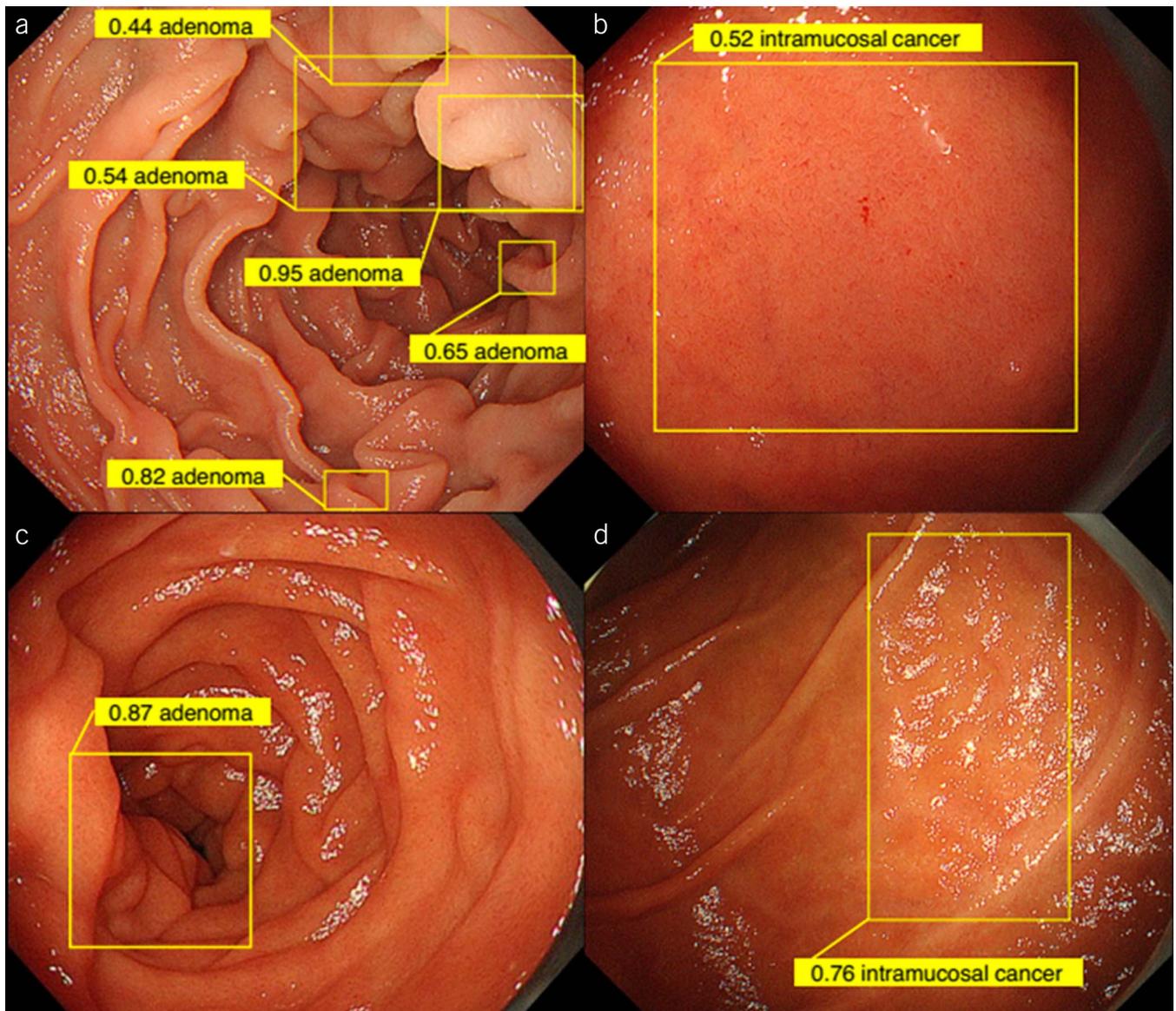

Figure 2. Most false-positives were caused by (**a**) normal duodenal folds, but some false-positives were caused by (**b**) normal duodenal mucosa, (**c**) duodenal papillary folds, and (**d**) low-quality images (e.g., halation).
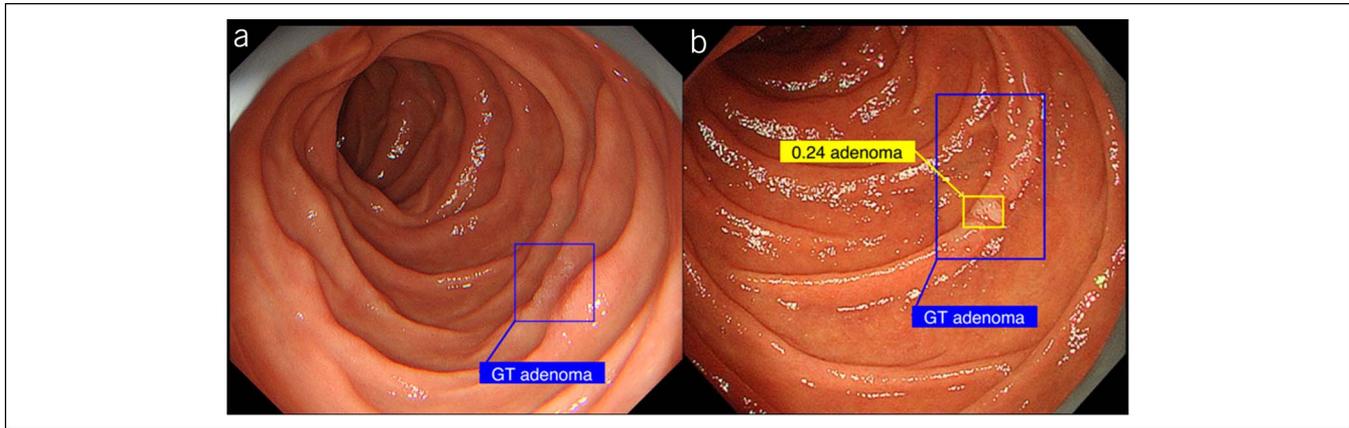
Figure 3. False-negative results were observed in 5.3% (21 of 399) of images. Most causes of false-negatives were lesions photographed from a distance, and even a skilled endoscopist had difficulty accurately detecting those lesions in those images only. (**a**) The CNN could not surround the lesion with a yellow frame. (**b**) The CNN was able to surround the lesion with a yellow frame, but the result was judged as negative because the cutoff score was set at 0.4. CNN, convolutional neural network; GT, ground truth.

Nevertheless, there is relatively little variation in the morphology and surface structure of SNADETs compared with other gastrointestinal tumors, so the detective performance of the CNN was very good with relatively few training images. Training and validation with a wider range of images obtained from other facilities and with other endoscopic devices are needed to further increase the versatility of the CNN. Second, this study used only high-resolution images; it is uncertain whether the CNN is capable of diagnosing SNADETs in low-quality images with halation, gradation of tone, blurring, or fouling by mucus or in low-resolution images acquired by transnasal endoscopy. For a wide application of this AI diagnostic technique to clinical examinations, the same detection rate must be maintained even with a low image quality. Robustness of the AI system can be obtained by using poor-quality images for the training process. However, including poor-quality images may impair the accuracy of the system. Therefore, robustness using poor-quality images should be considered after solidifying the performance of the system with a large amount of high-quality images. Third, the CNN was only trained with the images of SNADETs, and its performance was only verified with SNADETs and normal images. Application of this approach in general medical practice will require further training using benign lesions, such as Brunner's gland hyperplasia, lymphangiectasia, and duodenal papilla, and

submucosal tumor–like gastrointestinal stromal tumor because the future CNN will also learn submucosal tumor in the same manner by deep learning. Finally, this study did not compare the diagnostic performance of the system against an endoscopist, and whether these results are of sufficient sensitivity and specificity for introduction to clinical practice remains unknown.

We believe that these limitations can be resolved and that the AI-based diagnostic systems will soon allow for the real-time SNADET detection in everyday endoscopic practice. The technology will be required to analyze at least 30 images each second from the endoscopic video input. In the present study, all 399 images of the lesions were verified in 12 seconds. In other words, 33 images were analyzed each second, showing that this technical hurdle has been overcome. Adoption to actual clinical practice will make it possible to perform real-time image screening mechanically. With the exception of familial adenomatous polyposis, duodenal cancer typically affects people at around 70 years of age (21). The incidence of SNADETs is expected to increase as societal aging increases, and we can entrust the AI systems to provide increasing detection rates and diagnostic capacities for SNADETs.

In conclusion, an AI-based diagnostic system was developed based on thorough training of a CNN and displayed the ability to detect SNADETs with high sensitivity and specificity. We intend to conduct further studies that train the CNN with the images of normal and benign lesions and compare the diagnostic accuracy with endoscopy specialists with an overall aim of bringing this technology to everyday clinical practice.

Table 3. Comparison of diagnostic results between WLI and NBI

	AI diagnosis by WLI	AI diagnosis by NBI	*P* value
Sensitivity	92.9% (131/141)	98.5% (194/197)	0.009
Adenoma	89.9% (89/99)	98.7% (148/150)	0.002
HGD	100% (42/42)	100% (47/47)	1
Specificity	89.2% (511/573)	77.8% (84/108)	0.001

AI, artificial intelligence; HGD, high-grade dysplasia; NBI, narrow-band imaging; WLI, white-light imaging.
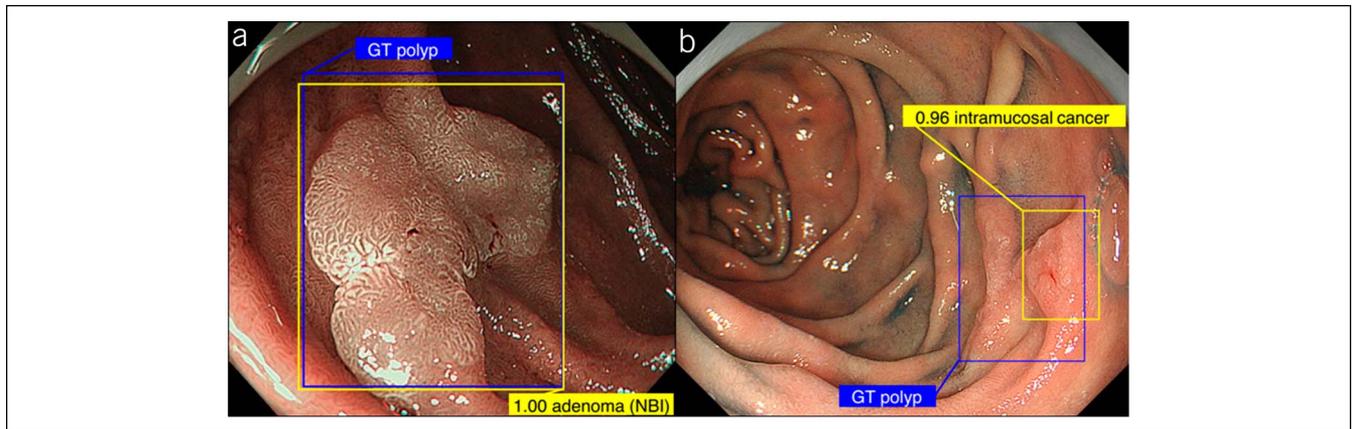
CONFLICTS OF INTEREST

ENDOSCOPY

Figure 4. (**a**) Narrow-band image and (**b**) indigo carmine image of the duodenal tumors detected by the CNN. CNN, convolutional neural network; GT, ground truth.

Study Highlights

WHAT IS KNOWN

✓ SNADETs rarely metastasize to lymph nodes, and most can be treated endoscopically.

✓ However, SNADETs are sometimes missed during esophagogastroduodenoscopic examination.

WHAT IS NEW HERE

✓ We used a novel algorithm to construct a CNN for detecting SNADETs. The CNN could detect all SNADETs in a short time.

TRANSLATIONAL IMPACT

✓ This system may be applicable in clinical practice to reduce the rate of missed SNADETs during esophagogastroduodenoscopy.

REFERENCES

1. Bilimoria KY, Bentrem DJ, Wayne JD, et al. Small bowel cancer in the United States: Changes in epidemiology, treatment, and survival over the last 20 years. Ann Surg 2009;249(1):63–71.
2. Jemal A, Siegel R, Ward E, et al. Cancer statistics, 2008. CA Cancer J Clin 2008;58(2):71–96.
3. Howe JR, Karnell LH, Scott-Conner C. Small bowel sarcoma: Analysis of survival from the national cancer data base. Ann Surg Oncol 2001;8(6):496–508.
4. Farley HA, Pommier RF. Surgical treatment of small bowel neuroendocrine tumors. Hematol Oncol Clin North Am 2016;30(1):49–61.
5. Ojha A, Zacherl J, Scheuba C, et al. Primary small bowel malignancies: Single-center results of three decades. J Clin Gastroenterol 2000;30(3):289–93.
6. Hochter W, Weingart J, Seib HJ, et al. Duodenal polyps. Incidence, histologic substrate and significance [in German]. Dtsch Med Wochenschr 1984;109(31-32):1183–6.
7. Jepsen JM, Persson M, Jakobsen NO, et al. Prospective study of prevalence and endoscopic and histopathologic characteristics of duodenal polyps in patients submitted to upper endoscopy. Scand J Gastroenterol 1994;29(6):483–7.
8. Murray MA, Zimmerman MJ, Ee HC. Sporadic duodenal adenoma is associated with colorectal neoplasia. Gut 2004;53:261–5.
9. Bibault JE, Giraud P, Burgun A. Big data and machine learning in radiation oncology: State of the art and future prospects. Cancer Lett 2016;382(1):110–7.
10. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542(7639):115–8.
11. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Jama 2016;316(22):2402–10.
12. Horie Y, Yoshio T, Aoyama K, et al. Diagnostic outcomes of esophageal cancer by artificial intelligence using convolutional neural networks. Gastrointest Endosc 2019;89(1):25–32.
13. Ohmori M, Ishihara R, Aoyama K, et al. Endoscopic detection and differentiation of esophageal lesions using a deep neural network. Gastrointest Endosc 2020;91(2):301–9.e1.
14. 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(3):407–14.
15. Shichijo S, Nomura S, Aoyama K, et al. Application of convolutional neural networks in the diagnosis of Helicobacter pylori infection based on endoscopic images. EBioMedicine 2017;25:106–11.
16. Shichijo S, Endo Y, Aoyama K, et al. Application of convolutional neural networks for evaluating Helicobacter pylori infection status on the basis of endoscopic images. Scand J Gastroenterol 2019;54:158–63.
17. 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(4):653–60.
18. Takiyama H, Ozawa T, Ishihara S, et al. Automatic anatomical classification of esophagogastroduodenoscopy images using deep convolutional neural networks. Sci Rep 2018;8(1):7497.
19. 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(2):416–21.e1.
20. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–44.
21. Thomas RM, Sobin LH. Gastrointestinal cancer. Cancer 1995;75(1 Suppl):154–70.