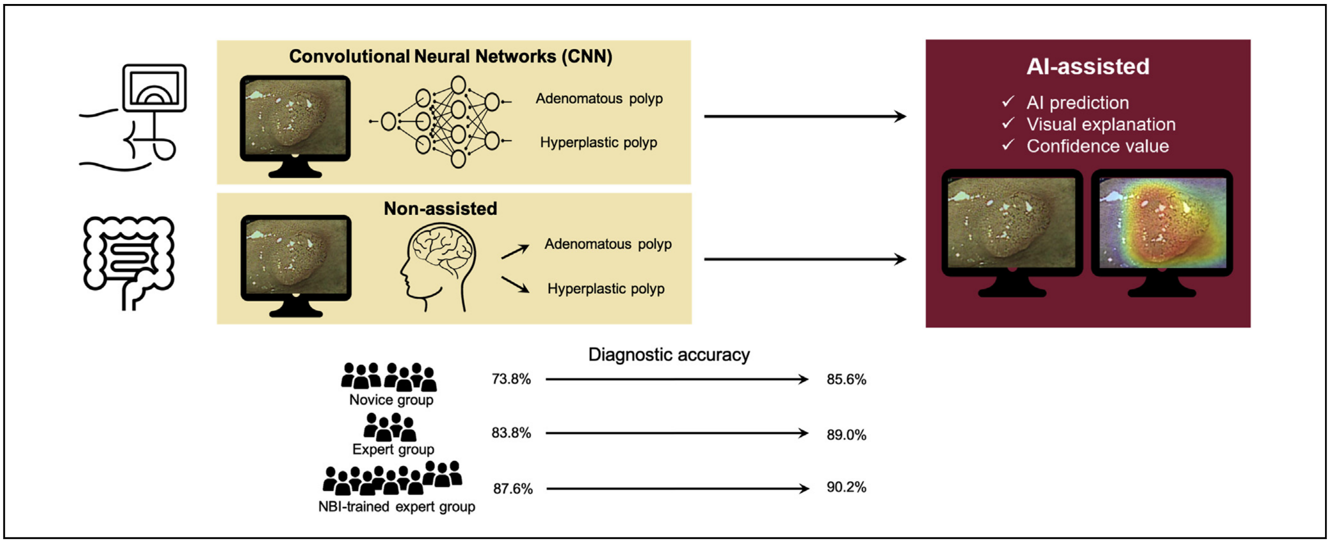
Improved Accuracy in Optical Diagnosis of Colorectal Polyps Using Convolutional Neural Networks with Visual Explanations

Eun Hyo Jin,[1,*] **Dongheon Lee**,[2,*] Jung Ho Bae,[1] Hae Yeon Kang,[1] Min-Sun Kwak,[1] Ji Yeon Seo,[1] Jong In Yang,[1] Sun Young Yang,[1] Seon Hee Lim,[1] Jeong Yoon Yim,[1] Joo Hyun Lim,[1] Goh Eun Chung,[1] Su Jin Chung,[1] Ji Min Choi,[1] Yoo Min Han,[1] Seung Joo Kang,[1] Jooyoung Lee,[3] Hee Chan Kim,[2,4,5] and Joo Sung Kim[1,3]

[1]Department of Internal Medicine, Healthcare Research Institute, Seoul National University Hospital Healthcare System Gangnam Center, Seoul, Korea; [2]Interdisciplinary Program in Bioengineering, Graduate School, Seoul National University, Seoul, Korea; [3]Department of Internal Medicine, Liver Research Institute, Seoul National University College of Medicine, Seoul, Korea; [4]Department of Biomedical Engineering College of Medicine, Seoul National University, Seoul, Korea; and [5]Institute of Medical & Biological Engineering, Medical Research Center, Seoul National University, Seoul, Korea



See editorial on page 2049.

BACKGROUND & AIMS: Narrow-band imaging (NBI) can be used to determine whether colorectal polyps are adenomatous or hyperplastic. We investigated whether an artificial intelligence (AI) system can increase the accuracy of characterizations of polyps by endoscopists of different skill levels. **METHODS:** We developed convolutional neural networks (CNNs) for evaluation of diminutive colorectal polyps, based on efficient neural architecture searches via parameter sharing with augmentation using NBIs of diminutive (≤5 mm) polyps, collected from October 2015 through October 2017 at the Seoul National University Hospital, Healthcare System Gangnam Center (training set). We trained the CNN using images from 1100 adenomatous polyps and 1050 hyperplastic polyps from 1379 patients. We then tested the system using 300 images of 180 adenomatous polyps and 120 hyperplastic polyps, obtained from January 2018 to May 2019. We compared the accuracy of 22 endoscopists of different skill levels (7 novices, 4 experts, and 11 NBI-trained experts) vs the CNN in evaluation of images (adenomatous vs hyperplastic) from 180 adenomatous and 120 hyperplastic polyps. The endoscopists then evaluated the polyp images with knowledge of the CNN-processed results. We conducted mixed-effect logistic and linear regression analyses to determine the effects of AI assistance on the accuracy of analysis of diminutive colorectal polyps by endoscopists (primary outcome). **RESULTS:** The CNN distinguished adenomatous vs hyperplastic diminutive polyps with 86.7% accuracy, based on histologic analysis as the reference standard. Endoscopists distinguished adenomatous vs hyperplastic diminutive polyps with 82.5% overall accuracy (novices, 73.8% accuracy; experts, 83.8% accuracy; and NBI-trained experts, 87.6% accuracy). With knowledge of the CNN-processed results, the overall accuracy of the endoscopists increased to 88.5% ($P < .05$). With knowledge of the CNN-processed results, the accuracy of novice endoscopists increased to 85.6% ($P < .05$). The CNN-processed results significantly reduced endoscopist time of diagnosis (from 3.92 to 3.37 seconds per polyp, $P = .042$). **CONCLUSIONS:** We developed a CNN that significantly increases the accuracy of evaluation of diminutive colorectal polyps (as adenomatous vs hyperplastic) and reduces the time of diagnosis by

CLINICAL AT

endoscopists. This AI assistance system significantly increased the accuracy of analysis by novice endoscopists, who achieved near-expert levels of accuracy without extra training. The CNN assistance system can reduce the skill-level dependence of endoscopists and costs.

Keywords: Deep Learning; Colorectal Cancer; Cancer Screening; Diagnostic.

WHAT YOU NEED TO KNOW

BACKGROUND AND CONTEXT

Narrow-band imaging can be used to determine whether colorectal polyps are adenomatous or hyperplastic. Artificial intelligence systems might increase the accuracy of characterization of narrow-band images of polyps by endoscopists.

NEW FINDINGS

The researchers developed a convolutional neural network system that significantly increased the accuracy of evaluation of narrow-band images of diminutive colorectal polyps (as adenomatous vs hyperplastic) and reduced time of diagnosis. This artificial intelligence assistance system also significantly increased the accuracy of analysis by novice endoscopists, who achieved near-expert levels of accuracy without extra training.

LIMITATIONS

This study was performed at a single center. Larger studies are needed.

IMPACT

The artificial intelligence system can increase the accuracy of evaluation of diminutive polyps by endoscopists and reduce time of evaluation and costs.

Colorectal cancer (CRC) is reported to be the third leading cause of death in the United States.[1] Furthermore, over the past several decades, the incidence of CRC has significantly increased in Asian countries, including Korea.[2] Most CRCs usually develop from preexisting adenomas, which are precancerous lesions, through the adenoma–carcinoma sequence.[3] In this regard, colonoscopy is currently the most important screening test for CRC because it can prevent CRC via the detection and subsequent removal of precancerous adenomatous polyps.[4] This is why adenoma detection is considered a key quality indicator of colonoscopy. Accordingly, considerable research efforts are directed toward the increase of the adenoma detection rate based on physician training and technical advances.

Although the detection and removal of adenoma contribute toward the reduction of CRC, the increased medical costs, including pathological analyses, also must be considered.[4] Most adenomatous polyps detected during screening colonoscopy are diminutive polyps (≤5 mm in size).[5,6] These rarely progress to CRC. However, the current practice is to subject all polyps to pathological evaluation.[5,6] Diminutive hyperplastic polyps of the rectosigmoid colon are very common benign lesions and do not require removal.[7] Moreover, discrepancies between endoscopic and pathologic diagnoses are not uncommon, and pathological diagnosis is not the gold standard for diagnosing colorectal polyps (≤3 mm).[8,9] Therefore, the application of an accurate endoscopic diagnosis before resection is advantageous because it prevents unnecessary resection and pathological evaluation. In this regard, optical diagnosis based on narrow-band imaging (NBI) can be used to predict the pathology of colorectal polyps and assist the distinction between adenomatous and hyperplastic colorectal polyps.[10] However, this implies that the endoscopist is sufficiently trained to perform adequate optical diagnosis.[11] Furthermore, such optical diagnosis is dependent on the endoscopist's skill and experience.[12] However, this limitation can be overcome with the newly developed computer-aided diagnosis (CADx).[13]

Meanwhile, recent advances in convolutional neural networks (CNN), one of the deep-learning approaches, have enabled their use in analyzing medical images. In this regard, many studies have reported on the convergence of the physician's skills and the use of artificial intelligence (AI) to afford accurate diagnoses.[14–19] In particular, in the optical diagnosis of colorectal polyps, CNN can afford high-performance diagnostics and detection from various colorectal-polyp images.[20–23] However, even if the AI approach affords high-performance colorectal-polyp diagnosis, endoscopists are currently required to perform a final diagnosis for the reasons of safety and accountability, and therefore, it is necessary to verify whether AI-based assistance can effectively aid in the final diagnosis.[17] Recently, Shahidi et al[9] introduced a real-time AI clinical decision support solution and showed that it could help the final diagnoses in the cases in which there were discrepancies between the endoscopic and pathologic diagnoses for diminutive polyps (≤3 mm).

In this study, we developed a deep-learning algorithm for the pathological classification of diminutive colorectal polyps based on NBI, and we compared its performance with those of endoscopists. Based on performance comparisons, we investigated the effect of AI assistance on the diagnostic accuracy of different skill-level groups of endoscopists to determine whether the polyps are adenomatous or hyperplastic from the NBI polyp images.

Methods

Study Design

This study was based on a multicenter study conducted from October 2015 to July 2019. It consisted of 3 stages: (1)

[*] Authors share co-first authorship.

Abbreviations: **AI, artificial intelligence; AutoML, automated machine learning; CADx, computer-aided diagnosis; CNN, convolutional neural networks; CRC, colorectal cancer; ENAS, efficient neural architecture search via parameter sharing; Grad-CAM, gradient-weighted class activation mapping; NBI, narrow-band images.**

Most current article

0016-5085/$36.00
https://doi.org/10.1053/j.gastro.2020.02.036

developed CNN for optical diagnosis of diminutive colorectal polyps, (2) conducted an endoscopic performance assessment and comparison with CNN (test 1), and (3) performed an endoscopic performance with knowledge of the CNN-processed results (test 2). The study protocol adhered to the ethical guidelines of the 1975 Declaration of Helsinki and its subsequent revisions, and was approved by the institutional review board (number H-1702-139-834). Written informed consent was obtained from all participating physicians.

Datasets

For the development of CNN for optical diagnosis, we retrospectively collected colonoscopic NBI of diminutive (≤5 mm) polyps from October 2015 to October 2017 at the Seoul National University Hospital, Healthcare System Gangnam. We used the routine pathology report to provide patient care. All polyps were removed using standard techniques and were subsequently evaluated by 1 of the 16 board-certified pathologists at the Seoul National University Hospital. We used an image set that was collected as part of the Gangnam-Real-Time Optical Diagnosis (READI) program as described in detail by Bae et al.[24] All colonoscopies were performed using high-definition colonoscopy (CF-HQ290; Olympus Co, Ltd., Tokyo, Japan) and acquired NBI with or without near-focus magnification. An endoscopist (E.H.J.) reviewed and selected well-focused, high-quality images with appropriate brightness values. If the optical diagnosis of a polyp was not compatible with the histological reports, the images were excluded. Finally, we trained the CNN with a total 1100 adenomatous polyps and 1050 hyperplastic polyps from 1379 patients (Supplementary Table 1). For the test dataset, we prospectively collected 300 polyp images (180 adenomatous polyps and 120 hyperplastic polyps) from January 2018 to May 2019 (Supplementary Table 2). Figure 1 shows the polyp samples presented in tests 1 and 2. The training, validation, and final test sets of endoscopic images of NBI polyps exhibited no overlap.

Development of CNNs

Preprocessing. The polyp regions-of-interest in the images were used for the training, and validations were conducted with the developed data acquisition program. These are described in Supplementary Figure 1 in detail. The shape of the polyp region-of-interest image was square and was resized to 128 × 128 to fit the input size of the CNN. A 5-fold cross-validation was applied as the training step, and the augmentation techniques were applied to generate the training datasets. A detailed description is presented in Supplementary Figure 2.

Search for CNN architecture. This study used an efficient neural architecture search via parameter sharing (ENAS), which is one of the automated machine learning (AutoML) methods.[25] The general process of training a standard CNN is limited in that it requires (1) specialized knowledge to design the architecture of CNN, and (2) trial-and-error experimentation to tune the hyperparameters that is time-consuming and expensive.[25] For this reason, AutoML has emerged and overcome the previous limitations and optimized both the network architecture and hyperparameters based on training methods. AutoML automates machine learning modeling, algorithmic selection, and hyperparameter tuning. Selecting and training CNN models requires the knowledge and experience of engineers and experimentation based on trial and errors. Therefore, the use of AutoML represents an attempt to optimize this complex and time-consuming process based on training, commonly referred to as the "learning to learn" methodology. Figure 2 represents the results of the search architecture based on training, and consists of repeating normal and reduction cells. A detailed description is presented in the "Description of CNN and Prediction Analysis" section in the supplementary material.

Training and utilization of searched CNNs. The training protocol of the model determined by the searching method is presented in the "Description of CNN and Prediction Analysis" section in the supplementary material. The utilization of the predicted results is as follows. The diagnostic confidence (probability) of hyperplastic and adenomatous polyps, which are the results of SoftMax in an inference step, were presented in a prospective study. In addition, a method of gradient-weighted class activation mapping (Grad-CAM)[26] was used to indicate the location of probabilistic evidence, and a heatmap overlaid on the polyp images diagnosed by the CNN was presented in a prospective study (Supplementary Figure 3). A detailed description is presented in the "Description of CNN and Prediction Analysis" section in the supplement.

This study compared the performance between inception-v3,[27] used in a previous study,[21,22] and the proposed method. Furthermore, we compared the results of the ENAS with those of the training set with the use of an augmentation method. The comparison of the performance outcomes include the accuracy, sensitivity, specificity, negative and positive predictive values, and diagnosis time, as listed in Table 1.

Evaluation of CNN and endoscopist performances. Twenty-two endoscopists participated in this study in 3 groups: (1) novices: 7 gastroenterology trainees with less than 2 years of colonoscopic experience from the Seoul National University Hospital, (2) experts: 4 board-certificated gastroenterologists with various experiences in NBI (co-authors: J.M.C., Y.M.H., S.J.K., J.L.), and (3) NBI-trained experts: 11 board-certificated gastroenterologists who were trained in optical diagnosis using NBIs (co-authors: J.H.B., H.Y.K., M.K., J.Y.S, J.I.Y., S.Y.Y., S.H.L., J.Y.Y., J.H.L., G.E.C., S.J.C.), commonly referred to as the Gangnam-READI program described in detail as in Bae et al[24] (Supplementary Table 3).

The following 2-stage tests were conducted based on the use of the validation dataset. All 300 NBI polyp images were de-identified and randomly ordered in each test. In test 1, each endoscopist independently evaluated the digital format of polyp NBIs to determine whether the polyp was adenomatous or hyperplastic test set on a retina display of a computer via an online survey. After a month, they performed test 2 in the same way as the previous test 1. In test 2, each endoscopist made an optical diagnosis based on the original polyp NBI (test 1) and the CNN-processed results. The AI results presented to the physician were as follows: (1) AI predicted the pathology (adenomatous or hyperplastic polyps), (2) confidence value, and (3) both original NBI polyp image and an explanation heatmap of the polyp NBI image obtained using Grad-CAM (Figure 1). In addition, each test also recorded the start and end times to calculate the average diagnostic time per polyp image. After 2 tests, we conducted individual surveys for the personality characteristics with the use of Grit-Original (Grit-O,

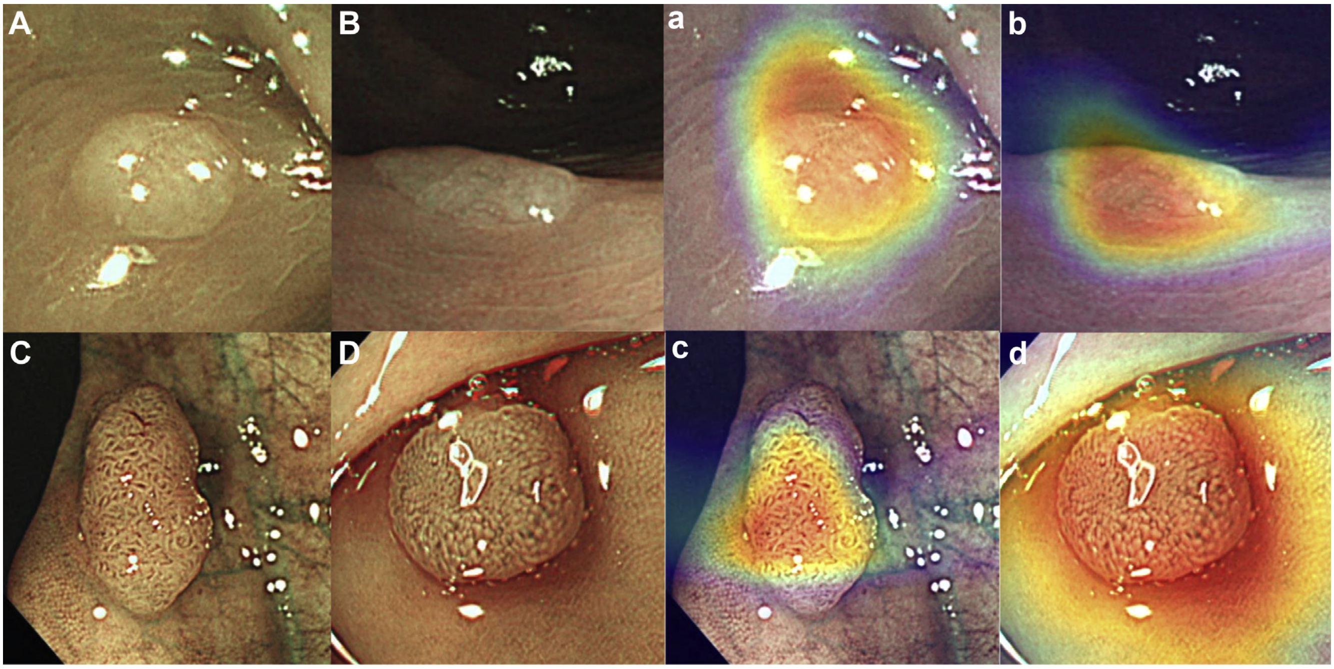


Figure 1. Illustration of experimental condition and polyp samples: (A, B, C, D) original NBIs, (a, b, c, d) visual explanation heatmap overlaid on original NBI. In test 1, we presented the original NBIs, and the original NBIs and visual explanation heatmap are presented in test 2.

Korean version) with 2 components, namely, consistency of interest and perseverance of effort.[28] The Grit-O was validated based on a questionnaire that comprised 12 items. It was scored on a 5-point scale (from 1 to 5). The summed score was divided by 12 to yield the final Grit score.[29]

Statistical Analyses

The main outcome of this study was to investigate the effect of AI assistance on the improvement of the optical diagnostic accuracy of endoscopists. The optical-diagnosis performances of the CNN and the endoscopists (test 1) and those of the endoscopists with AI assistance (test 2) were evaluated and compared with the use of the McNemar test. We developed a mixed-effects logistic regression model to estimate the effect of AI assistance on the subgroups. Wilcoxon signed rank tests were used to assess differences of diagnostic time between nonassisted and AI-assisted assessments. The Grit scores were analyzed using correlations and linear regression analyses. For all the tests, a *P* value of .05 was considered to indicate statistical significance, and a *P*-value correction was performed. All calculations were performed with the SAS (version 9.3; SAS Institute, Cary, NC) software package.

Results

Performance Comparison Between Endoscopists and CNN Performance (Test 1)

In this study, the CNN selected using ENAS with augmentation techniques exhibited an optical diagnostic accuracy of 86.7% (95% confidence interval 82.3–90.3), with a sensitivity of 83.3% and a specificity of 91.7%. The diagnostic performance of the CNN was compared with those of 22 endoscopists (Table 2). Five of the 7 novices yielded significantly lower diagnostic accuracies (47.7%–79.0%) than the CNN ($P < .05$). Only 1 endoscopist (E1, 77.3%) of the 4 experts demonstrated significantly lower diagnostic accuracy than the CNN ($P < .05$). Among the 11 NBI-trained expert endoscopists, 1 endoscopist (N-TE4, 92.7%) demonstrated statistically higher diagnostic accuracy than the CNN ($P = .011$).

Diagnostic Accuracy Improvement With AI Assistance (Test 2)

The overall accuracy of optical diagnosis was significantly increased with the use of AI assistance (82.5% to 88.5%, $P < .05$) (Supplementary Table 4). Although AI assistance appeared to improve the endoscopist performances, it must be considered that this increase can vary according to the endoscopist experiences. In the novice group, all endoscopists demonstrated performances with significantly increased accuracies ($P < .05$), and 4 of them demonstrated performances with greater accuracy than the algorithm. In the expert group, 2 endoscopists demonstrated performances with significantly improved accuracies (E1, $P = .01$; E4, $P = .001$), and 1 (E4) achieved higher accuracy than the algorithm. In the NBI-trained expert group, 3 endoscopists (N-TE1, N-TE2, N-TE11) demonstrated performances with significantly improved accuracies ($P < .05$). Interestingly, 1 endoscopist (N-TE2) was already more accurate than the algorithm without AI assistance.

AI assistance and endoscopic experience. The optical-diagnosis performances of the novices, expert endoscopists, and NBI-trained expert endoscopists were 73.8%, 83.8%, and 87.6%, respectively, and their diagnostic

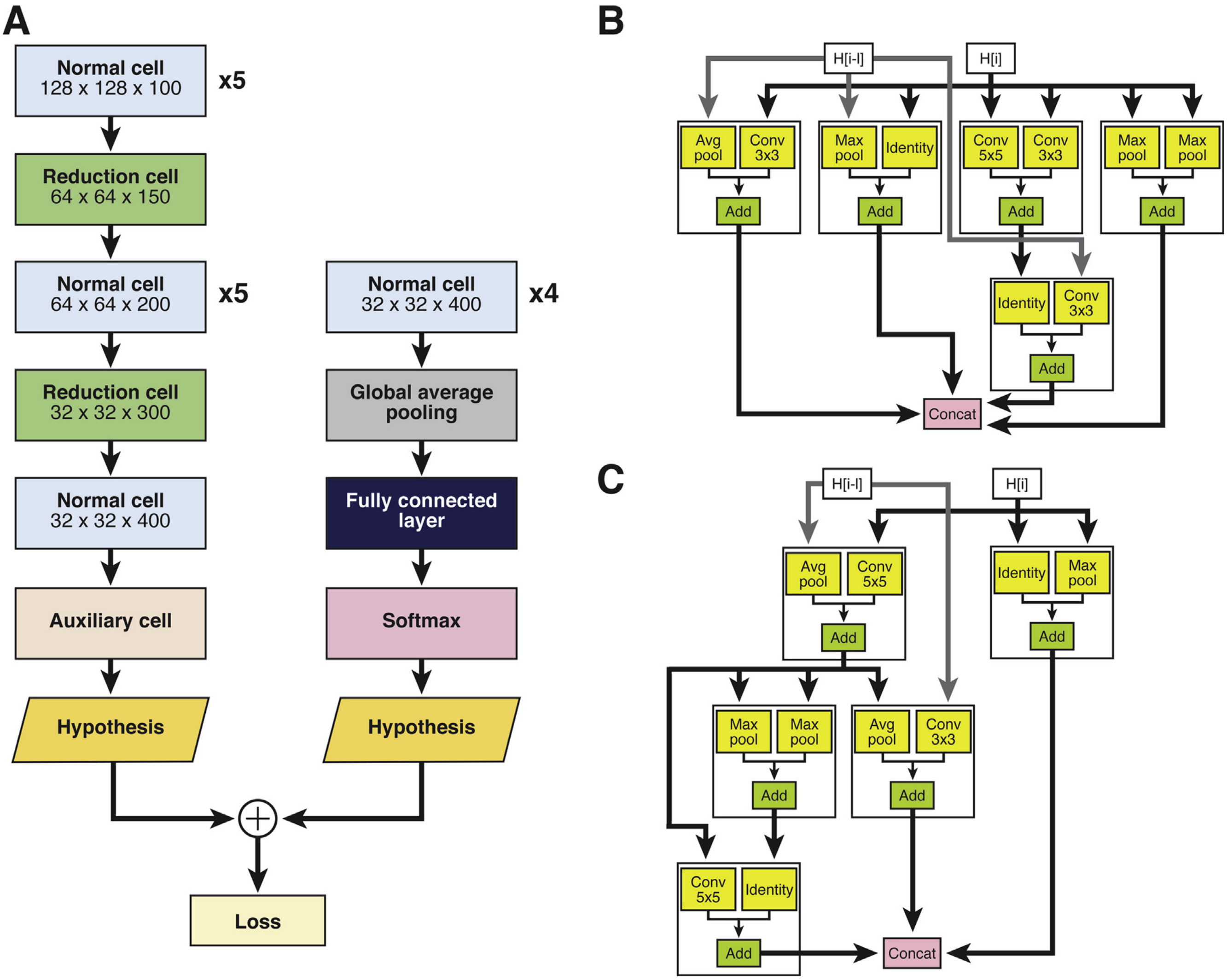


Figure 2. Architecture of the CNNs for the classification of NBI of polyps searched based on the method of neural architecture search. (*A*) Full architecture of CNNs searched by the proposed method, (*B*) architecture of normal cell, and (*C*) architecture of a reduction cell.

accuracy improved with AI assistance (85.6%, 89.0%, 90.2%, respectively, Figure 3). Without AI assistance (test 1), the novice group demonstrated a significantly lower accuracy than both the experts ($P = .049$) and the NBI-trained experts ($P = .001$). With AI assistance (test 2), the accuracy of the novices significantly improved, and there was no statistical difference when performances were compared with those of the expert group ($P = .102$) (Supplementary Figure 4). A detailed results of sensitivity and specificity were presented in Supplementary Figure 5.

AI assistance and diagnostic time. The average time for the AI algorithm to diagnose each polyp was 0.01 second, which is significantly shorter than the time taken by the endoscopists (Table 3). Herein, we note that AI assistance offered an interpretable explanation such that endoscopists can diagnose faster. In particular, the diagnostic time per polyp reduced from 4.44 to 3.68 seconds in the case of the NBI-trained expert group ($P = .033$).

Personality traits and acceptance of AI assistance. The acceptance of AI assistance by the endoscopist also forms an important factor in diagnosis. This acceptance factor can be reflected by the personality trait of the grit. The traits of the grit are defined as the perseverance and passion for long-term goals, and they reflect the ability of an individual to sustain long-term efforts and overcome obstacles in realizing goals.[30] In our study, the mean participant grit score was 3.56 (Table 4). Overall, we observed a moderate correlation between grit and AI-assisted diagnostic accuracy ($r = 0.51$, $P = .015$) (Supplementary Figure 6). Conversely, there was no correlation between grit and diagnostic accuracy without AI assistance.

Discussion

In this study, we investigated the effect of AI assistance on 22 endoscopists in accurately predicting the pathology of polyp NBI. We found that AI assistance with an interpretable explanation could improve both the optical diagnostic accuracy and diagnostic speed regardless of endoscopic

Table 1. The CNN Performance Comparison Between a Previous Method and Proposed Methods

	Accuracy, n (%)	Sensitivity, n (%)	Specificity, n (%)	Positive predictive value, n (%)	Negative predictive value, n (%)	Diagnostic time (s)
Inception-v3	245/300 (81.67)	144/180 (80)	101/120 (84.17)	141/160 (88.34)	103/140 (73.72)	8.42/300
ENAS*	256/300 (85.33)	147/180 (81.67)	109/120 (90.83)	145/160 (90.83)	107/140 (76.76)	3.62/300
ENAS* + augmentation	260/300 (86.7)	150/180 (83.3)	110/120 (91.7)	150/160 (93.8)	110/140 (78.6)	

ENAS, efficient neural networks architecture search via parameter sharing.

experience. The diagnostic accuracy increased maximally in the novice group, and it was not significantly different from that of the expert group. Herein, we note that AI assistance can aid even well-trained expert endoscopists in increasing their diagnostic accuracies and reduce the diagnosis duration. In this section, we discuss the results considering the various aspects of the study.

In this study, we used NBI for polyp classifications. NBI was acquired with the use of a standard colonoscope, and not a magnifying endoscope or endocytoscope. Here, we note that recently, CADx systems have demonstrated satisfactory diagnostic capability in predicting the histology based on images captured with a magnification endoscope (×80) and endocytoscope (×500).[22,31] However, these advanced imaging modalities are not commonly used in clinical practice. In our case, even without magnification, the obtained accuracy was only slightly lower (86.7%) than that of previous studies that used magnification (88.0%, 90.1%).[22,31] Meanwhile, we note that although a large number of adenomatous polyps exhibited high-grade dysplasia (71/284, 25.0%) in a previous study,[22] all adenomatous polyps exhibited low-grade dysplasia in our study. Herein, we note that diminutive colorectal polyps smaller than 5 mm mostly exhibit low-grade dysplasia, whereas high-grade dysplasia has been reported in only 0.3% to 1.2% of cases.[5,32]

Recent advances in AI technology have accelerated the development of CADx[14,17] toward the distinction between

Table 2. Diagnostic Accuracy Stratified Based on the Viewing Condition (Nonassisted vs AI-assisted)

	Non-assisted (T1)				AI-assisted (T2)			
	Accuracy			T1 vs AI	Accuracy			T1 vs T2
Observer	n	Percent	95% CI	(*P* value)	n	Percent	95% CI	(*P* value)
AI	260/300	86.7	(82.3–90.3)					
Novice (n = 7)								
N1	236/300	78.7	(73.6–83.2)	.009	264/300	88.0	(83.8–91.5)	<.0001
N2	237/300	79.0	(73.9–83.5)	.003	261/300	87.0	(82.7–90.6)	.001
N3	245/300	81.7	(76.8–85.9)	.075	262/300	87.3	(83–90.9)	.024
N4	255/300	85.0	(80.4–88.8)	.522	269/300	89.7	(85.7–92.9)	.035
N5	226/300	75.3	(70.1–80.1)	<.0001	247/300	82.3	(77.5–86.5)	.007
N6	143/300	47.7	(41.9–53.5)	<.0001	237/300	79.0	(73.9–83.5)	<.0001
N7	207/300	69.0	(63.4–74.2)	<.0001	258/300	86.0	(81.6–89.7)	<.0001
Expert endoscopist (n = 4)								
E1	232/300	77.3	(72.2–81.9)	.002	259/300	86.3	(81.9–90)	.001
E2	254/300	84.7	(80.1–88.6)	.460	263/300	87.7	(83.4–91.2)	.208
E3	265/300	88.3	(84.1–91.7)	.515	270/300	90.0	(86–93.2)	.466
E4	254/300	84.7	(80.1–88.6)	.439	276/300	92.0	(88.3–94.8)	.001
NBI-trained expert endoscopist (n = 11)								
N-TE1	258/300	86.0	(81.6–89.7)	.808	276/300	92.0	(88.3–94.8)	.011
N-TE2	264/300	88.0	(83.8–91.5)	.593	282/300	94.0	(90.7–96.4)	.004
N-TE3	267/300	89.0	(84.9–92.3)	.362	265/300	88.3	(84.1–91.7)	.746
N-TE4	278/300	92.7	(89.1–95.3)	.011	282/300	94.0	(90.7–96.4)	.371
N-TE5	269/300	89.7	(85.7–92.9)	.225	271/300	90.3	(86.4–93.4)	.683
N-TE6	264/300	88.0	(83.8–91.5)	.617	268/300	89.3	(85.3–92.6)	.505
N-TE7	270/300	90.0	(86–93.2)	.181	280/300	93.3	(89.9–95.9)	.059
N-TE8	263/300	87.7	(83.4–91.2)	.714	266/300	88.7	(84.5–92)	.602
N-TE9	256/300	85.3	(80.8–89.1)	.537	256/300	85.3	(80.8–89.1)	1.000
N-TE10	250/300	83.3	(78.6–87.4)	.211	259/300	86.3	(81.9–90)	.150
N-TE11	252/300	84.0	(79.4–88)	.339	270/300	90.0	(86–93.2)	.011

CI, confidence interval; T1, test 1; T2, test 2.

Table 3. Comparison of Average Diagnostic Times for Each Polyp Image Between CNN and Endoscopists

	Diagnostic time per polyp (s)		
	Nonassisted (T1)	AI-assisted (T2)	*P* value
CNN	0.01	0.01	1.000
Overall	3.92	3.37	.042
Novice	3.24	3.18	.866
Expert	3.67	2.84	.068
NBI-trained expert	4.44	3.68	.033

T1, test 1; T2, test 2.

adenomatous and hyperplastic colorectal polyps.[21,22] Previous studies have reported on classification systems that have demonstrated expert-endoscopist–level accuracy of optical diagnosis.[22,33] However, the effects of these models applied to endoscopists are not well understood. Our study is the first attempt to identify how the diagnostic capabilities of endoscopists differ between both AI-unassisted (test 1) and AI-assisted diagnoses (test 2). Our study showed that AI assistance augments the physician's judgment, thereby improving the accuracy of optical diagnosis and the shortening of the diagnostic time.

Unlike the general method of training standard CNN, we used an ENAS, which is one of the AutoML methods.[25] Previous CNN medical imaging studies had been selected and trained the defined CNN models, such as inception-v3,[27] which yielded high-performance outcomes in the ImageNet competition.[34] However, these CNN architectures performed tasks on general datasets, and not on specific datasets, such as the NBI polyp. Thus, we used the proposed method to search the CNN architecture by training that was optimized for polyp NBI. In addition, the proposed method is faster in formulating a diagnosis compared with previous studies given that it is based on a better graphics processing unit performance, smaller batch size, and smaller training image size. Accordingly, it is considered to be suitable for real-time diagnosis. We also found that the diagnostic performances of the ENAS with the augmentation techniques for the flat polyp cases were improved compared with the single-ENAS method in conjunction with the endoscopist diagnoses.[35,36] Considering that AI did not recognize this type of polyp well in previous studies,[37] the use of the proposed methods confirmed that the combination of various augmentation techniques could compensate for the lack of training data and improve the performance. As a result of this process, the loss graph of the training and validation sets in Supplementary Figure 7 indicates that the 2 decreasing loss patterns are the same. Given that there is no significant difference between the 2 indicates that overfitting is minimized.

In the application of medical AI technology, it is important that physicians can understand the AI results to accept AI. Here, we mention that deep-learning methods are "black boxes" because it is impossible to explain why the AI arrived at a specific decision.[38] In this context, we note that recently AI explanation methods have been developed to enable humans to comprehend how the AI predictions are made.[19,39] In this study, we presented the AI results to physicians in the following manner: AI-predicted pathology with confidence value and both original polyp NBI and NBI with generated heatmaps using Grad-CAM methods.[26] We visualized the highlights that overlaid the polyp NBI for predicted evidence. This interpretable explanation of AI results can aid the endoscopist to accept AI assistance, thereby contributing to increased diagnostic accuracy.

In the study, we divided the endoscopists into 3 groups, novices, experts, and NBI-trained experts, based on their skill. Novice users were gastroenterology fellows with no experience in NBI polyp diagnosis. The mean durations (± standard deviation) of colonoscopy experiences were 5.5 ± 3.1 years in the case of the expert group, and 13.0 ± 5.6 years in the case of the NBI-trained group. Both experts and NBI-trained experts were board-certified gastroenterologists and they performed more than 600 colonoscopies per year. Expert groups had various experiences with NBI, and 3 of them had minor experiences with NBI. By contrast, the NBI-trained expert group participants were trained in optical diagnosis using NBI for 1 year and group performance of participating endoscopists met or surpassed the Preservation and Incorporation of Valuable Endoscopic Innovations threshold.

We found that AI assistance is most effective in aiding novices rather than experts. All the novices demonstrated significantly increased diagnostic accuracy, and their results were not inferior to those of experts. It should be noted that in previous studies, the "nonexpert" results showed marked interobserver variability, and these nonexperts could not achieve acceptable accuracy in the optical diagnosis of diminutive polyps with NBI.[40,41] To overcome this limitation, many researchers have attempted to develop AI diagnostic algorithms that can allow nonexperts to demonstrate improved accuracy of optical diagnosis as a clinical tool.[42–44] Our study demonstrates that AI assistance may aid in augmenting the abilities of nonexperts with limited training in optical diagnosis to take better decisions.

In our study, the 11 NBI-trained experts participated in a training program for optical diagnosis using NBI from September 2015 to September 2016.[24] This NBI-trained expert group demonstrated the highest accuracy of 87.6%, which is thought to be attributed to the effects of training. Even in the case of NBI-trained experts whose performance was better than AI algorithms, the accuracy increased and the diagnosis time reduced with AI assistance. These results suggest that AI assistance can also be useful for experts in actual clinical situations.

Grit is a positive, noncognitive personality trait characterized by the ability to persevere during difficulties combined with powerful motivation to achieve a goal.[45] Grit has been found to be a superior predictor of success in high-achievement fields.[29] Higher grit has been found to correlate with higher performance in medical school, whereas lower grit has been found to correlate with increased surgical residency training drop-out rates.[46,47] Previous studies have shown that doctors exhibit an average grit score in the

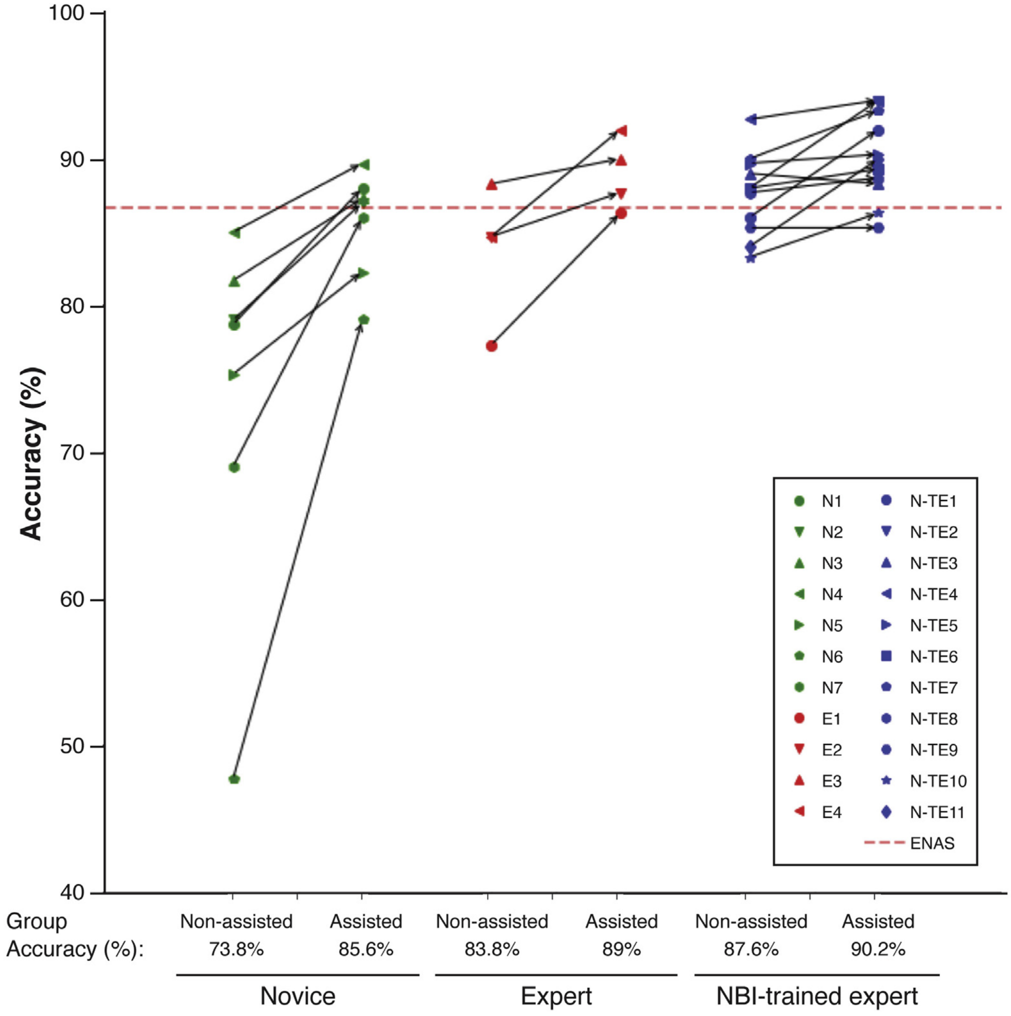


Figure 3. Improved accuracy of optical diagnosis with AI assistance classified by the group.

range of 3.5 to 3.7.[30,45] In our case, the participating endoscopists exhibited an average grit score of 3.56. Our study findings show that endoscopists with high grit scores could flexibly accept AI assistance, thereby increasing the diagnostic accuracy. In this study, we found that high grit, particularly in terms of the consistency of interest, correlated with high accuracy, which translated to a passion to achieve and maintain strong motivation for overcoming obstacles. This result indicates the possibility that certain personality traits of the endoscopist can affect the acceptance of AI technology.

This study has several limitations. First, we developed a CNN based on high-quality images. However, in clinical practice, the acquired images may be of poor quality, such as out-of-focus or blurred images. Second, this study does not focus on real-time optical diagnoses. We

Table 4. Mean Score for Grit (5-Scale) and Strength of Correlation Between Optical Diagnostic Accuracy (r = Correlation Coefficient)

				Optical diagnostic accuracy			
	Overall			Nonassisted (T1)		AI-assisted (T2)	
	Mean	SD	IQR	Correlation, *r*	*P* value	Correlation, *r*	*P* value
Grit score	3.561	0.47	3.22–3.92	0.3	.1768	0.51	.0148
Consistency of interest	3.386	0.59	3.00–3.83	0.38	.0799	0.56	.0069
Perseverance of effort	3.735	0.5	3.50–4.00	0.11	.6175	0.31	.1651

SD, standard deviation; IQR, interquartile range.

considered only 2 in vitro tests to compare the performances of endoscopists with and without AI assistance. We cropped and resized images to fit the CNN's input size. These hand-crafted, extracted images could be different from actual colonoscopy images. In actual colonoscopy, the endoscopist could observe polyps at various angles and in continuous frames to predict pathology. Endoscopic video streams could be more useful than still images.[21] Third, our training and test datasets consisted of tubular adenoma with low-grade and hyperplastic polyps. We excluded diminutive polyps with serrated lesions, and other benign conditions, such as inflammatory polyps or lymphoid follicles. Further studies are needed on other types of colorectal polyps with various pathological findings. Fourth, the confidence value, the probabilistic diagnosis of CNN, is not always reliable because the diagnosis is not based on the same approach as that used for humans.[38] Therefore, to solve the uncertainty issue,[43] the Bayesian deep-learning method has been studied that can be trained with weights of probability distribution rather than with the use of fixed-weight CNN values.[48] Finally, because Grad-CAM is a technology that was applied independently on the proposed CNN architecture and on the training methods, the presented heatmap results were not stable because the results were different at each CNN layer.

In conclusion, AI assistance is useful for the improvement of the accuracy of the optical diagnosis of diminutive polyps and for the achievement of shorter diagnostic times. In particular, we found that AI assistance was most effective for novices because they could achieve accuracies similar to those of experts without training or effort. In this manner, by reducing the diagnostic-capability differences between physicians, pathological examinations can be replaced by accurate optical diagnoses with AI assistance that can contribute to significant reductions of medical costs.

Supplementary Material

Note: To access the supplementary material accompanying this article, visit the online version of *Gastroenterology* at www.gastrojournal.org, and at https://doi.org/10.1053/j.gastro.2020.02.036.

References

1. Cronin KA, Lake AJ, Scott S, et al. Annual report to the nation on the status of cancer, part i: national cancer statistics. Cancer 2018;124:2785–2800.
2. Sung JJ, Lau JY, Goh KL, et al. Increasing incidence of colorectal cancer in Asia: implications for screening. Lancet Oncol 2005;6:871–876.
3. Leslie A, Carey FA, Pratt NR, et al. The colorectal adenoma-carcinoma sequence. Br J Surg 2002;89:845–860.
4. Corley DA, Jensen CD, Marks AR, et al. Adenoma detection rate and risk of colorectal cancer and death. N Engl J Med 2014;370:1298–1306.
5. Ponugoti PL, Cummings OW, Rex DK. Risk of cancer in small and diminutive colorectal polyps. Dig Liver Dis 2017;49:34–37.
6. Lieberman D, Moravec M, Holub J, et al. Polyp size and advanced histology in patients undergoing colonoscopy screening: implications for CT colonography. Gastroenterology 2008;135:1100–1105.
7. Rex DK. Narrow-band imaging without optical magnification for histologic analysis of colorectal polyps. Gastroenterology 2009;136:1174–1181.
8. Ponugoti P, Rastogi A, Kaltenbach T, et al. Disagreement between high confidence endoscopic adenoma prediction and histopathological diagnosis in colonic lesions </= 3 mm in size. Endoscopy 2019;51:221–226.
9. Shahidi N, Rex DK, Kaltenbach T, et al. Use of endoscopic impression, artificial intelligence, and pathologist interpretation to resolve discrepancies from endoscopy and pathology analyses of diminutive colorectal polyps. Gastroenterology 2020;158:783–785.e1.
10. Hewett DG, Kaltenbach T, Sano Y, et al. Validation of a simple classification system for endoscopic diagnosis of small colorectal polyps using narrow-band imaging. Gastroenterology 2012;143:599–607.e1.
11. Ignjatovic A, Thomas-Gibson S, East JE, et al. Development and validation of a training module on the use of narrow-band imaging in differentiation of small adenomas from hyperplastic colorectal polyps. Gastrointest Endosc 2011;73:128–133.
12. Vleugels JLA, Dijkgraaf MGW, Hazewinkel Y, et al. Effects of training and feedback on accuracy of predicting rectosigmoid neoplastic lesions and selection of surveillance intervals by endoscopists performing optical diagnosis of diminutive polyps. Gastroenterology 2018;154:1682–1693.e1.
13. Misawa M, Kudo SE, Mori Y, et al. Characterization of colorectal lesions using a computer-aided diagnostic system for narrow-band imaging endocytoscopy. Gastroenterology 2016;150:1531–1532.e3.
14. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. Nat Med 2019;25:24–29.
15. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88.
16. Shen D, Wu G, Suk H-I. Deep learning in medical image analysis. Annu Rev Biomed Eng 2017;19:221–248.
17. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25:44–56.
18. Mansour RF. Deep-learning-based automatic computer-aided diagnosis system for diabetic retinopathy. Biomed Eng Lett 2018;8:41–57.
19. Sayres R, Taly A, Rahimy E, et al. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. Ophthalmology 2019;126:552–564.
20. Billah M, Waheed S. Gastrointestinal polyp detection in endoscopic images using an improved feature extraction method. Biomed Eng Lett 2018;8:69–75.
21. Byrne MF, Chapados N, Soudan F, et al. Real-time differentiation of adenomatous and hyperplastic diminutive

colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. Gut 2019;68:94–100.

22. Chen P-J, Lin M-C, Lai M-J, et al. Accurate classification of diminutive colorectal polyps using computer-aided analysis. Gastroenterology 2018;154:568–575.
23. Wang P, Xiao X, Brown JRG, et al. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. Nat Biomed Eng 2018; 2:741.
24. Bae JH, Lee C, Kang HY, et al. Improved real-time optical diagnosis of colorectal polyps following a comprehensive training program. Clin Gastroenterol Hepatol 2019;17:2479–2488.e4.
25. Pham H, Guan MY, Zoph B, et al. Efficient neural architecture search via parameter sharing. arXiv preprint 2018;arXiv:1802.03268.
26. Selvaraju RR, Cogswell M, Das A, et al. Grad-cam: visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision, Venice, 2017, pp. 618–626.
27. Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, 2016, pp. 2818–2826.
28. Ung Lee S-WL, Young-Chul Shin, Dong-Won Shin, et al. Reliability and validity of Korean version of GRIT. Anxiety and Mood 2019;15:53–60.
29. Duckworth AL, Peterson C, Matthews MD, et al. Grit: perseverance and passion for long-term goals. J Pers Soc Psychol 2007;92:1087–1101.
30. Dam A, Perera T, Jones M, et al. The relationship between grit, burnout, and well-being in emergency medicine residents. AEM Educ Train 2019;3:14–19.
31. Mori Y, Kudo SE, Chiu PW, et al. Impact of an automated system for endocytoscopic diagnosis of small colorectal lesions: an international web-based study. Endoscopy 2016;48:1110–1118.
32. Chaput U, Alberto SF, Terris B, et al. Risk factors for advanced adenomas amongst small and diminutive colorectal polyps: a prospective monocenter study. Dig Liver Dis 2011;43:609–612.
33. Mori Y, Kudo SE, Berzin TM, et al. Computer-aided diagnosis for colonoscopy. Endoscopy 2017;49:813–819.
34. Deng J, Dong W, Socher R, et al. Imagenet:. a large-scale hierarchical image database. IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, 2009:248–255.
35. Wang J, Perez L. The effectiveness of data augmentation in image classification using deep learning. Convolutional Neural Networks Vis. Recognit 2017;11.
36. Castro E, Cardoso JS, Pereira JC. Elastic deformations for data augmentation in breast cancer mass detection. IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), Las Vegas, NV, 2018, pp. 230–234.
37. 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.e8.
38. Castelvecchi D. Can we open the black box of AI? Nature 2016;538:20–23.
39. Poplin R, Varadarajan AV, Blumer K, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng 2018;2:158–164.
40. Kuiper T, Marsman WA, Jansen JM, et al. Accuracy for optical diagnosis of small colorectal polyps in nonacademic settings. Clin Gastroenterol Hepatol 2012; 10:1016–1020; quiz e79.
41. Rogart JN, Jain D, Siddiqui UD, et al. Narrow-band imaging without high magnification to differentiate polyps during real-time colonoscopy: improvement with experience. Gastrointest Endosc 2008;68: 1136–1145.
42. Adebayo J, Gilmer J, Muelly M, et al. Sanity checks for saliency maps. In: Advances in Neural Information Processing Systems, 2018;9505-9515.
43. Begoli E, Bhattacharya T, Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. Nat Mach Intell 2019;1:20–23.
44. Wagner J, Kohler JM, Gindele T, et al. Interpretable and fine-grained visual explanations for convolutional neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, 2019;9097-9107.
45. Halliday L, Walker A, Vig S, et al. Grit and burnout in UK doctors: a cross-sectional study across specialties and stages of training. Postgrad Med J 2017;93:389–394.
46. Miller-Matero LR, Martinez S, MacLean L, et al. Grit: a predictor of medical student performance. Educ Health (Abingdon) 2018;31:109–113.
47. Salles A, Lin D, Liebert C, et al. Grit as a predictor of risk of attrition in surgical residency. Am J Surg 2017; 213:288–291.
48. Kendall A, Gal Y. What uncertainties do we need in bayesian deep learning for computer vision? Advances in neural information processing systems, 2017;5574-5584.

Author names in bold designate shared co-first authorship.

Received November 19, 2019. Accepted February 20, 2020.

Address correspondence to: Joo Sung Kim, MD, PhD, Department of Internal Medicine and Liver Research Institute, Seoul National University College of Medicine, 101 Daehak-ro, Jongno-gu, Seoul 03080, Republic of Korea. e-mail: jooskim@snu.ac.kr; fax: +82-2112-5635; or Hee Chan Kim, PhD, Department of Biomedical Engineering, Seoul National University Hospital, 101 Daehak-ro, Jongno-gu, Seoul 03080, Republic of Korea. e-mail: hckim@snu.ac.kr; fax: +82-2-745-7870.

Acknowledgments
We thank Changhyun Lee, Jinju Choi, Jue Lie Kim, Ayoung Lee, Seona Park, Hosim Soh, and Seung Wook Hong for their support on this project.

CRediT Authorship Contributions
Eun Hyo Jin, MD, PhD (Conceptualization: Equal; Formal analysis: Lead; Methodology:

Lead; Writing – original draft: Equal). Dongheon Lee, PhD student (Conceptualization: Equal; Formal analysis: Supporting; Methodology: Supporting; Software: Lead; Writing – original draft: Equal). Jung Ho Bae, MD (Methodology: Equal). Hae Yeon Kang, MD, PhD (Data curation: Equal). Min-Sun Kwak, MD, PhD (Data curation: Equal). Ji Yeon Seo, MD, PhD (Data curation: Supporting). Jong In Yang, MD, PhD (Data curation: Equal). Sun Young Yang, MD, PhD (Data curation: Equal). Seon Hee Lim, MD, PhD

(Data curation: Equal). Jeong Yoon Yim, MD, PhD (Data curation: Equal). Joo Hyun Lim, MD, PhD (Data curation: Equal). Goh Eun Chung, MD, PhD (Data curation: Equal). Su Jin Chung, MD, PhD (Data curation: Equal). Ji Min Choi, MD, PhD (Data curation: Equal). Yoo Min Han, MD (Data curation: Equal). Seung Joo Kang, MD, PhD (Data curation: Equal). Jooyoung Lee, MD (Data curation: Equal). Hee Chan Kim, PhD (Software: Lead; Supervision: Equal; Writing – review & editing: Equal). Joo-Sung Kim, MD, PhD (Investigation: Lead; Supervision: Lead; Writing – review & editing: Equal).

Conflict of interest
The authors disclose no conflicts.

Funding
EHJ was supported by the Ministry of Science and ICT (MSIT), South Korea, under the Information Technology Research Center (ITRC) support program (IITP-2019-2018-0-01833) supervised by the Institute for Information & Communications Technology Promotion (IITP).

Supplementary Material

Dataset Acquisition and Augmentation Methods

Dataset Acquisition

A dataset acquisition program was developed for region-of-interest (ROI) analyses of polyp image acquisitions from original polyp images. This program provides various functionalities, including the ability to import images in a folder, draw ROIs with the mouse, and save the coordinates of the polyp in the image. The program was developed in MATLAB (MATLAB R2017a; MathWorks Inc., Natick, MA), as shown in Supplementary Figure 1. In the data acquisition step, the NBI has a size of 1280 × 960 pixels (200%), and the polyp region is cropped within a selected ROI.

Dataset Augmentation Techniques

As part of the training of the CNNs, the augmentation technique is used to improve performance. In this experiment, the number of training sets was increased 5 times based on the application of the augmentation techniques, and yielded the highest performance based on several experiments. The applied methods were a combination of linear transformations (zoom: 0.15, shear: 0.3, rotation: 60°) and an elastic transformation[1] (σ: 12, random 3 × 3 grid) using the software packages OpenCV (version 3.4.1) and elasticdeform (version 0.4.6). The results of the augmentation techniques are shown in Supplementary Figure 2.

Description of CNN and Prediction Analysis

Efficient Neural Architecture Search via Parameter Sharing

Efficient neural architecture search via parameter sharing is one of the AutoML methods that uses recurrent neural networks (RNN)[2] and reinforcement learning (RL)[3] methods to determine the architecture of the deep-learning model. In this case, the RNN that determines the architecture of the model is called a controller, and the model created by the controller is called a child network. The controller used the RL method to yield a higher child network performance based on the accuracy of the generated child network. In turn, the child network trained each sampled child network with a general image training method and with the use of a training dataset.

The proposed method is the architecture searching method and the procedure is as follows.

1. The controller RNN generates hyperparameters for the architectural design of convolutional neural networks.
2. As the controller RNN constructs the architecture, it calculates the accuracy of the validation set based on training until the loss converges.
3. To maximize the expected validation accuracy of the constructed architecture, a policy gradient method is used to optimize the hyperparameters of the controller RNN.
4. This process is repeated to search for the optimal architecture design.

Specifically, this study used a micro search to design small modules and then connected them to CNN.[4] The modules consisted of normal cells and reduction cells, and these 2 modules formed the networks in a repeating architecture.

In addition, 5 types of operations were determined within the modules based on training, and the types were (1) identity, (2) separable convolution with kernel sizes of 3 × 3 and 5 × 5, and (3) average pooling and max pooling with a kernel size of 3 × 3. The hyperparameters used for the training of the controller RNN and micro search were determined based on experiments as follows. The RNN controller learning rate was 0.003, the child learning rate was 0.0005 to 0.05, the L2 regularization was 1e-4, and the numbers of the child layer, branches, and child cells were 5, 5, and 15, respectively.

The hardware development environment included the NVIDIA Titan V, graphics processing unit, and the software was Python (version 3.4.2; Python Software Foundation, Beaverton, OR), TensorFlow (version 1.11.0; Google, Mountain View, CA). It was developed with reference to https://github.com/melodyguan/enas/.

CNN Training

The training protocol of the model determined by the searching method is as follows. The model was trained with an epoch of 450 and with a batch size of 10. An Adam optimizer[5] was used with a learning rate of 0.0001. In addition, a weighted cross-entropy method[6] was used to solve a class imbalance issue, and the ratio of the training datasets was not precisely 1:1.

Grad-CAM as a Basis for Diagnosis

Grad-CAM is one of the explainable AI techniques that presents the results of the CNN as a probabilistic representation of a heatmap overlaid on an image.[7] The closer the color of the heatmap is to blue, the lower is the probability, and the closer the color is to red, the higher is the probability. The results of the applied Grad-CAM are shown in Supplementary Figure 3.

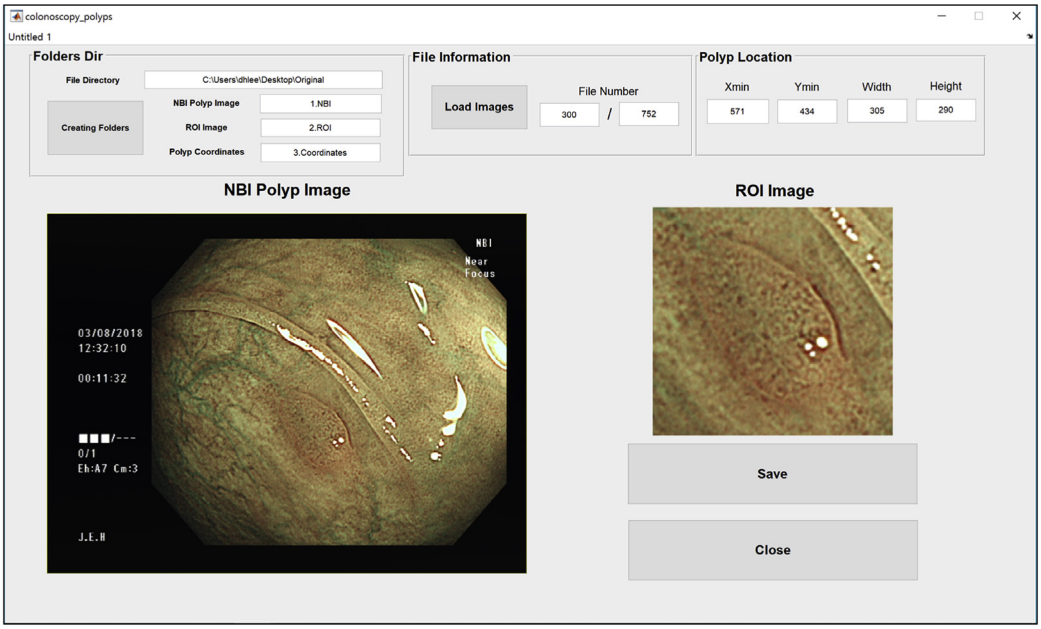
t-Stochastic Neighbor Embedding

t-Distributed stochastic neighbor embedding (t-SNE) is a dimension reduction method, whereby high-dimensional data are embedded as low-dimensional data and are visualized.[8] We defined the similarity between the data in a high-dimensional space represented by probability values and the similarity between the data in an embedding (low-dimensional) space. Accordingly, the gradient descent was used so that the difference between the 2 similarities was small.

In this study, features of the validation set were extracted from the last layer of the trained CNN. The number of features was 1024, and the features of the last layer reduced the dimension to 2, with a learning rate of 200, and with 1000 iterations based on the use of the package scikit-learn machine learning package (version 0.19.1; https://scikit-learn.org). The results are shown in Supplementary Figure 8.

Supplementary References

1. Castro E, Cardoso JS, Pereira JC. Elastic deformations for data augmentation i breast cancer mass detection. IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), Las Vegas, NV, 2018, pp. 230–234.
2. Zaremba W, Sutskever I, Vinyals O. Recurrent neural network regularization. arXiv preprint 2014;arXiv:1409.2329.
3. Sutton RS, Barto AG. Introduction to reinforcement learning. Cambridge: MIT Press;1998.
4. Zoph B, Vasudevan V, Shlens J, et al. Learning transferable architectures for scalable image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018:8697–8710.
5. Kingma DP, Adam BJ. A method for stochastic optimization. arXiv preprint 2014;arXiv:1412.6980.
6. Aurelio YS, de Almeida GM, de Castro CL, et al. Learning from imbalanced data sets with weighted cross-entropy function. Neural Processing Letters 2019;50:1937–1949.
7. Selvaraju RR, Cogswell M, Das A, et al. Grad-cam: visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision, Venice, 2017, pp. 618–626.
8. Maaten Lvd, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9:2579–2605.



Folders Dir

File Directory: C:\Users\dhlee\Desktop\Original

NBI Polyp Image: 1.NBI

ROI Image: 2.ROI

Polyp Coordinates: 3.Coordinates

File Information

File Number: 300 / 752

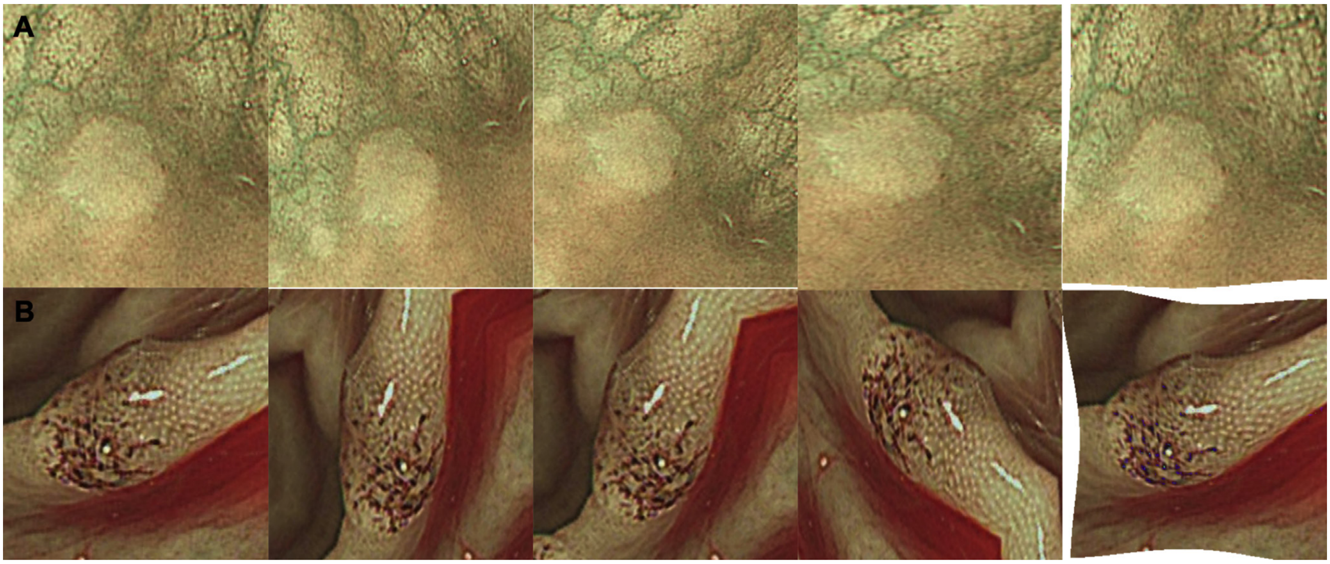
Polyp Location

Xmin	Ymin	Width	Height
571	434	305	290

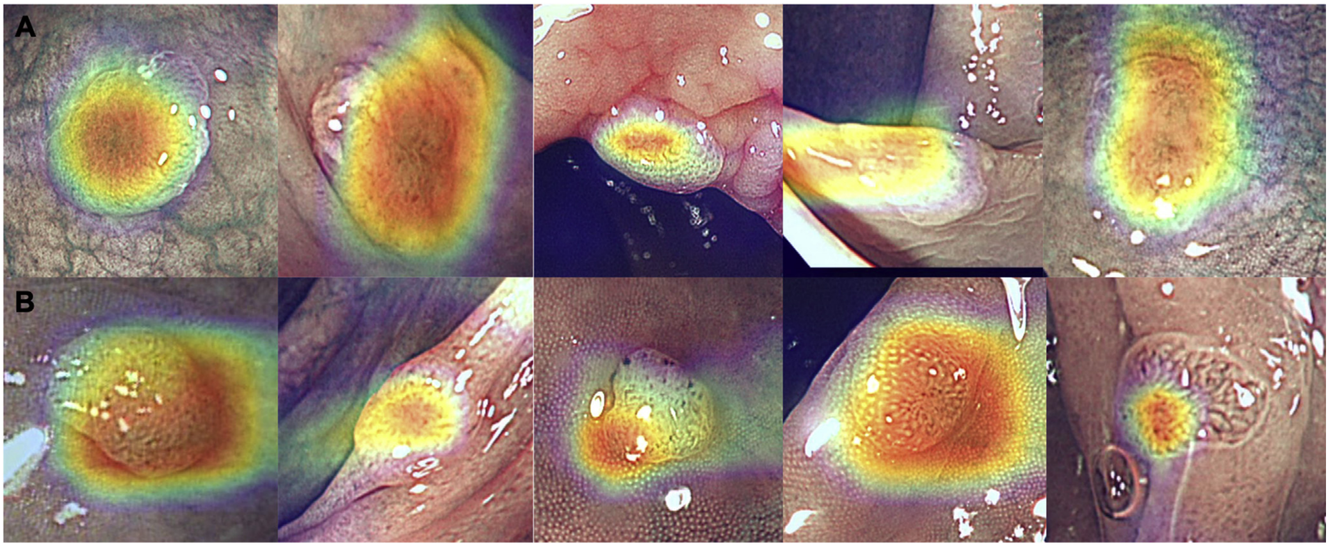
NBI Polyp Image

ROI Image

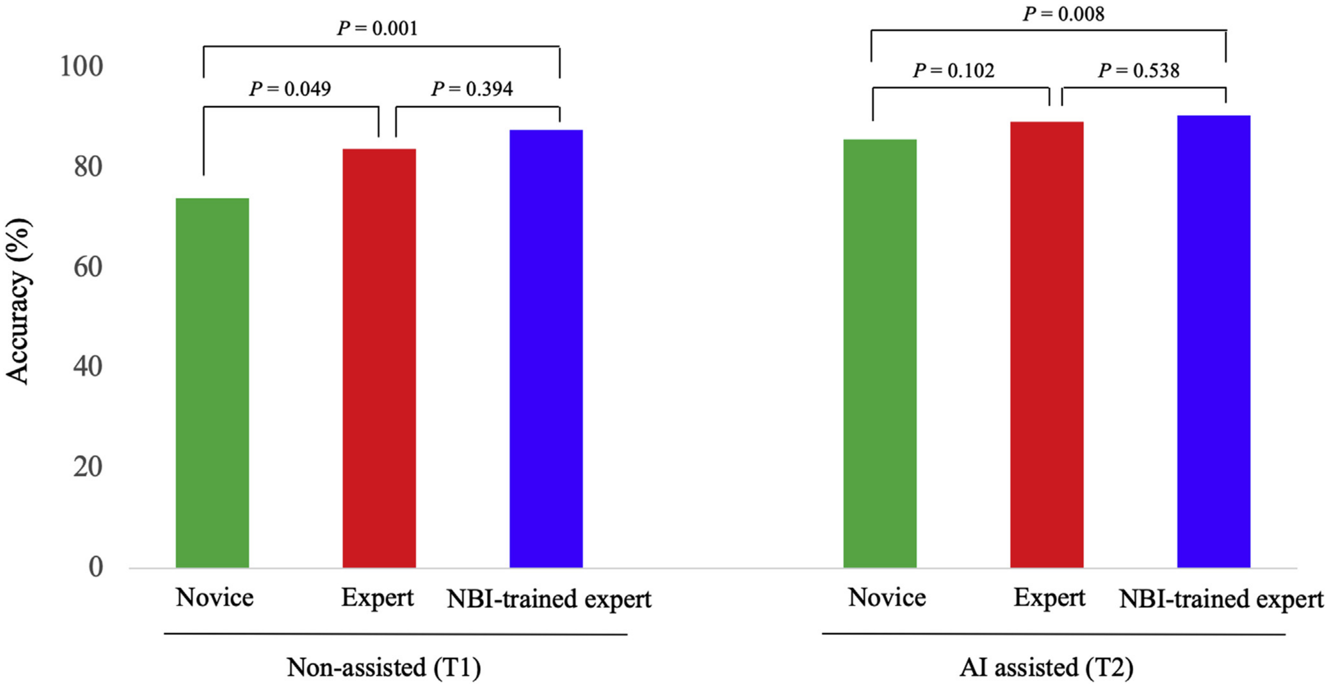
Supplementary Figure 1. Data acquisition program. The program provides functionalities for loading original polyp NBIs, selection, and saving.



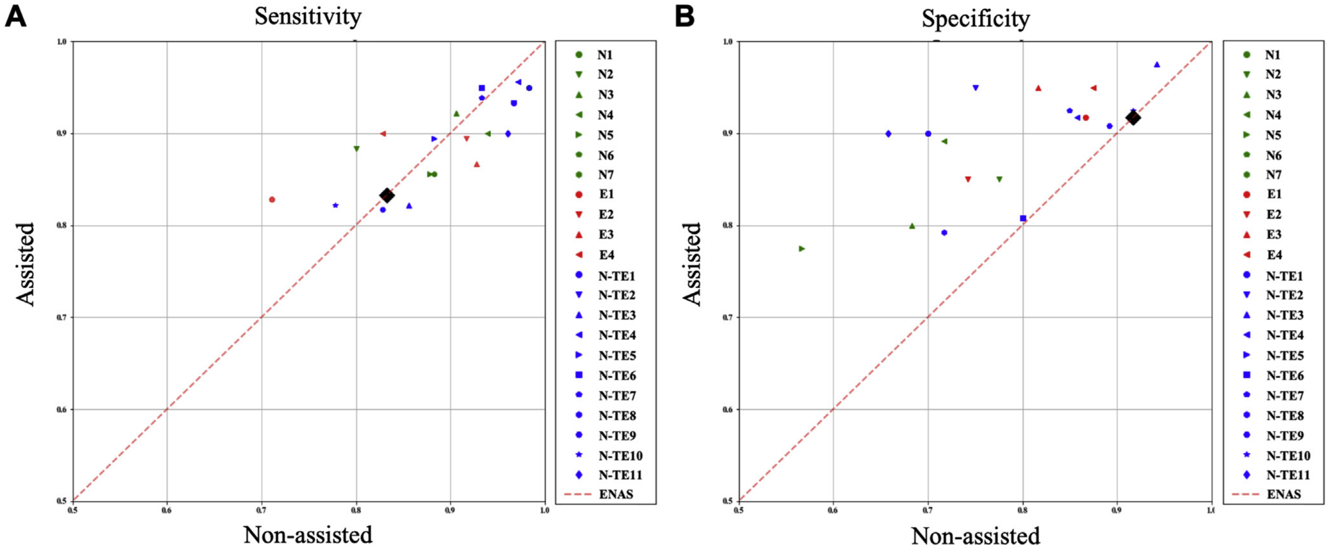
Supplementary Figure 2. Applied augmentation techniques. (*A*) Augmentation results of hyperplastic polyp images, and (*B*) augmentation results of adenomatous polyp images.



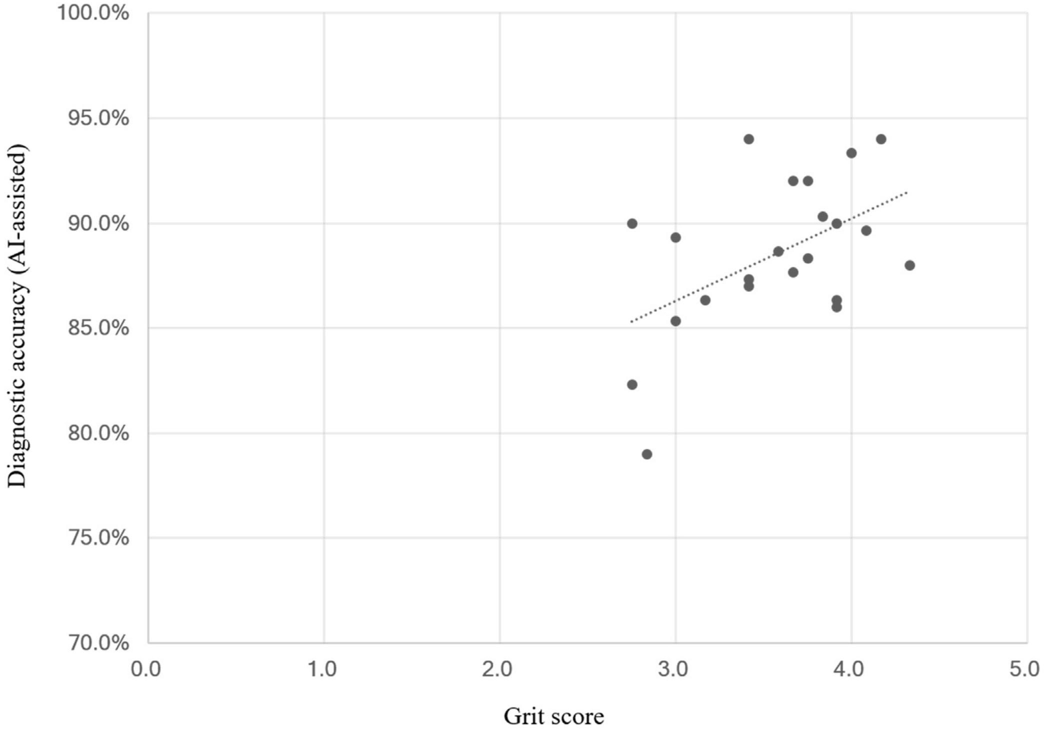
Supplementary Figure 3. Results of probabilistic diagnosis as a heatmap on polyp images using Grad-CAM. (*A*) Heatmap results overlaid on hyperplastic polyp images. (*B*) Heatmap results on adenomatous polyp images.



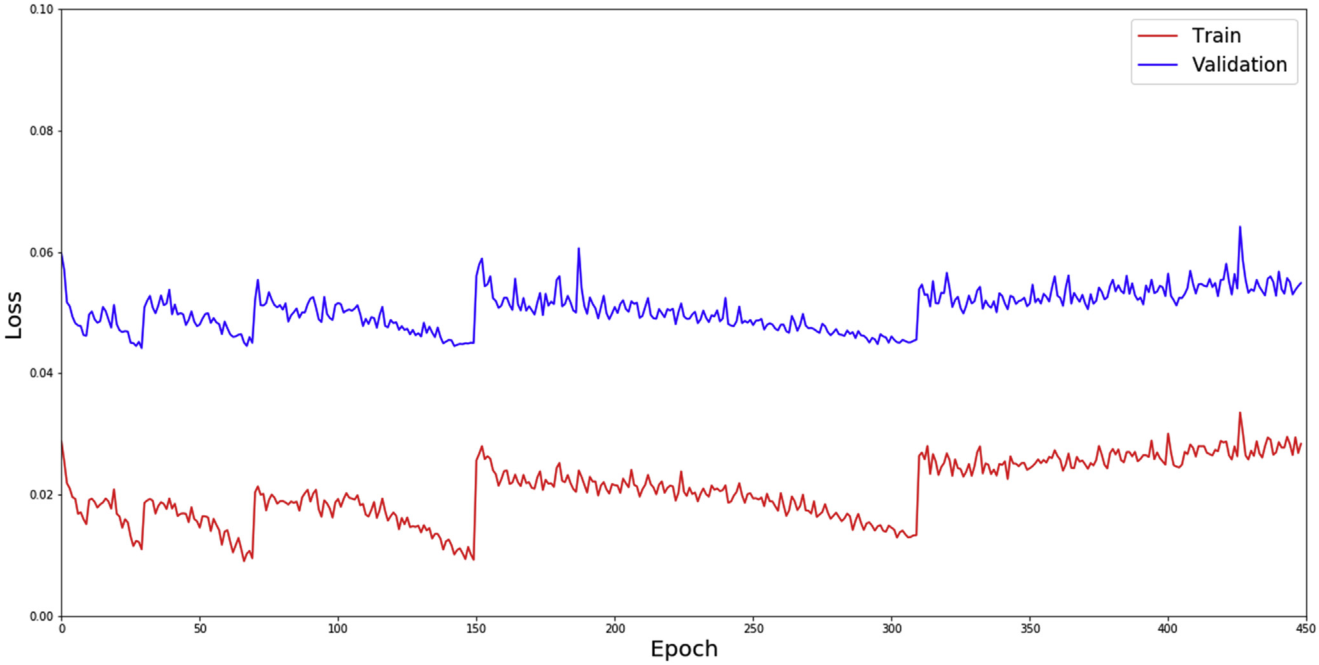
Supplementary Figure 4. Comparisons of the diagnostic accuracy outcomes according to endoscopic experiences in non-assisted and AI-assisted conditions.



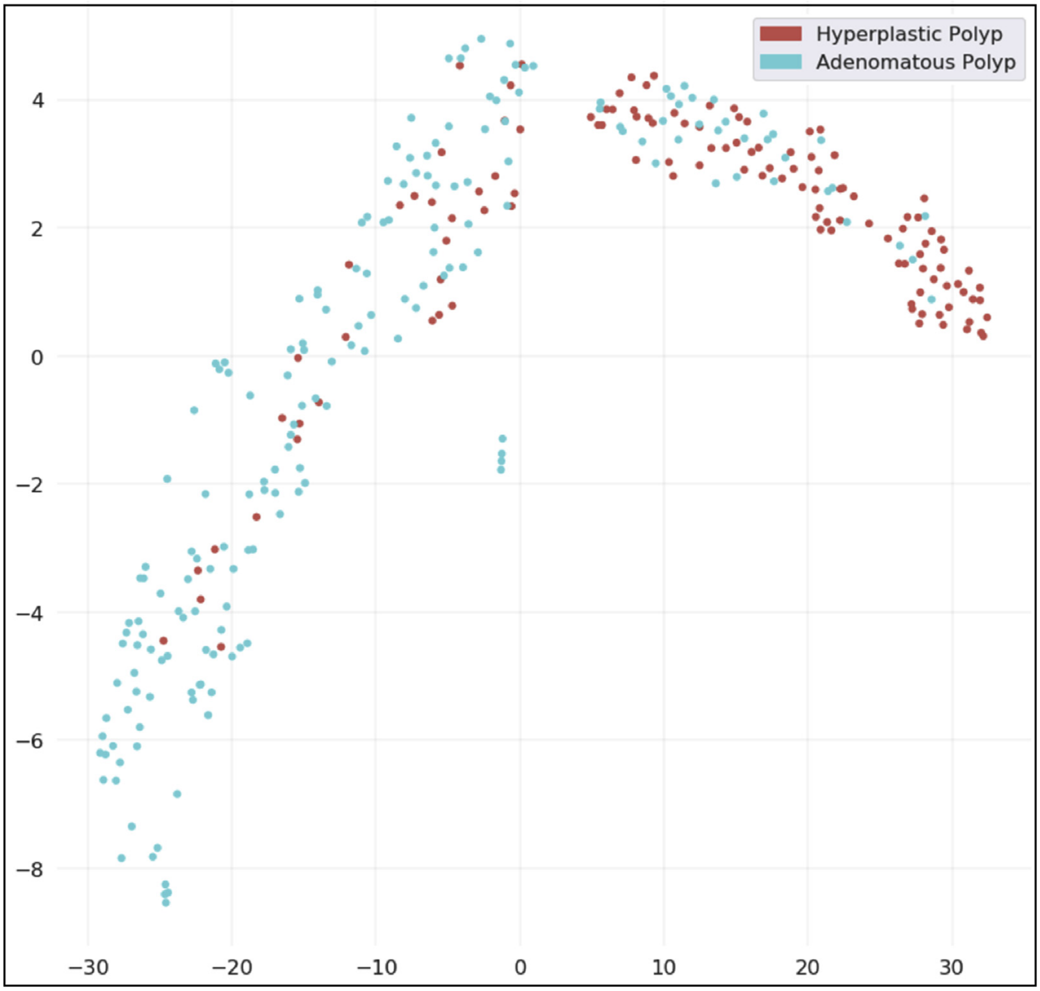
Supplementary Figure 5. Scatterplots of (*A*) sensitivity, and (*B*) specificity for each AI-assisted condition (*y*-axis) compared with nonassisted condition (*x*-axis) for participating endoscopists. Results show that AI assistance increased specificity.



Supplementary Figure 6. Scatter plot for AI-assisted optical diagnosis against grit score ($r = 0.51$, $P = .015$).



Supplementary Figure 7. Loss graph of training and validation sets.



Supplementary Figure 8. Result following the application of t-stochastic neighbor embedding to NBI polyp images.

Supplementary Table 1. Polyp Characteristics of Training Set (N = 2150)

	Adenomatous polyp (n = 1100)	Hyperplastic polyp (n = 1050)	*P*
Location			< .0001
- Ascending colon	362 (32.9)	179 (17.0)	
- Transverse colon	310 (28.2)	171 (16.3)	
- Descending colon	119 (10.8)	55 (5.2)	
- Rectosigmoid colon	309 (28.1)	645 (61.4)	
Using NF view			< .0001
- without NF view	96 (8.7)	171 (16.3)	
- with NF view	1004 (91.3)	879 (83.7)	
Gross			< .0001
- IIa	499 (45.4)	894 (85.1)	
- Is	505 (45.9)	152 (14.5)	
- Isp	96 (8.7)	4 (0.4)	

NOTE. Values are n (%).
NF, near focus.

Supplementary Table 2. Patient Information and Polyp Characteristics in the Validation Test Set (N = 300)

	Adenomatous polyp (n = 180)	Hyperplastic polyp (n = 120)	*P*
Sex			.062
- Male	127 (70.6)	97 (80.8)	
- Female	53 (29.4)	23 (19.2)	
Age (mean ± SD)	60.0 ± 10.0	54.9 ± 9.9	.000
Location			.000
- Ascending colon	61 (33.9)	26 (21.7)	
- Transverse colon	61 (33.9)	15 (12.5)	
- Descending colon	14 (7.8)	13 (10.8)	
- Rectosigmoid colon	44 (24.4)	66 (55.0)	
Using NF view			.752
- without NF view	12 (6.7)	10 (8.3)	
- with NF view	168 (93.3)	110 (91.7)	
Gross			.002
- IIa (flat)	131 (72.8)	106 (88.3)	
- Is (sessile)	34 (18.9)	13 (10.8)	
- Isp (subpedunculated)	15 (8.3)	1 (0.8)	

NOTE. Values are n (%).
NF, near focus; SD, standard deviation.

Supplementary Table 3. Baseline Characteristics of Participating Endoscopists (N = 22)

	n (%)
Sex	
Male	4 (18.2)
Female	18 (81.8)
Colonoscopy experience (*y*)	
<2	7 (31.8)
2–9	6 (27.3)
10–14	5 (22.7)
≥15	4 (18.2)
Estimated cumulative colonoscopy volume	
<1000	4 (18.2)
1000–2500	4 (18.2)
2500–4999	5 (22.7)
5000–9999	6 (27.3)
≥10,000	3 (13.6)
Observed polyp with NBI mode in usual practice	
Not at all	1 (4.5)
>25%	4 (18.2)
>50%	5 (22.7)
>75%	6 (27.3)
All	6 (27.3)
Usefulness of NBI mode for optical diagnosis	
Not at all	0 (0.0)
>25%	3 (13.6)
>50%	7 (31.8)
>75%	7 (31.8)
All	5 (22.7)

Supplementary Table 4. Comparison of the Diagnostic Accuracy According to Endoscopic Experiences in Nonassisted and AI-assisted Cases

	Nonassisted (T1)		AI-assisted (T2)		T1 vs T2				
	Accuracy (%)	SE	Accuracy (%)	SE	Difference	Lower	Upper	SE	*P*
Group									
Novice	73.8	2.86	85.6	1.19	11.86	7.27	16.45	2.19	<.0001
Expert	83.8	3.78	89.0	1.57	5.25	−0.82	11.32	2.90	.0861
NBI-trained expert	87.6	2.28	90.2	0.95	2.55	−1.11	6.21	1.75	.1619
Overall	82.5	1.61	88.5	0.67	6.00	3.41	8.59	1.24	.0001

SE, standard error.