Thieme

Automated classification of gastric neoplasms in endoscopic images using a convolutional neural network

Authors

Bum-Joo Cho[1,2,3], Chang Seok Bang[3,4,5], Se Woo Park[4,5], Young Joo Yang[3,4,5], Seung In Seo[4,5], Hyun Lim[4,5], Woon Geon Shin[4,5], Ji Taek Hong[4,5], Yong Tak Yoo[6], Seok Hwan Hong[6], Jae Ho Choi[3], Jae Jun Lee[3,7], Gwang Ho Baik[4,5]

Institutions

1 Department of Ophthalmology, Hallym University College of Medicine, Chuncheon, Korea
2 Interdisciplinary Program in Medical Informatics, Seoul National University College of Medicine, Seoul, Korea
3 Institute of New Frontier Research, Hallym University College of Medicine, Chuncheon, Korea
4 Department of Internal Medicine, Hallym University College of Medicine, Chuncheon, Korea
5 Institute for Liver and Digestive Diseases, Hallym University, Chuncheon, Korea
6 Dudaji Inc., Seoul, Korea
7 Department of Anesthesiology and Pain medicine, Hallym University College of Medicine, Chuncheon, Korea

submitted 27.11.2018
accepted after revision 19.6.2019

Bibliography
DOI https://doi.org/10.1055/a-0981-6133
Published online: 2019 | Endoscopy

ISSN 0013-726X

Corresponding author
Chang Seok Bang, MD, PhD, Department of Internal Medicine, Hallym University College of Medicine, Sakju-ro 77, Chuncheon, Gangwon-do 24253, South Korea
Fax: +82-33-2418064
csbang@hallym.ac.kr

Supplementary material
Online content viewable at:
https://doi.org/10.1055/a-0981-6133

ABSTRACT

Background Visual inspection, lesion detection, and differentiation between malignant and benign features are key aspects of an endoscopist's role. The use of machine learning for the recognition and differentiation of images has been increasingly adopted in clinical practice. This study aimed to establish convolutional neural network (CNN) models to automatically classify gastric neoplasms based on endoscopic images.

Methods Endoscopic white-light images of pathologically confirmed gastric lesions were collected and classified into five categories: advanced gastric cancer, early gastric cancer, high grade dysplasia, low grade dysplasia, and non-neoplasm. Three pretrained CNN models were fine-tuned using a training dataset. The classifying performance of the models was evaluated using a test dataset and a prospective validation dataset.

Results A total of 5017 images were collected from 1269 patients, among which 812 images from 212 patients were used as the test dataset. An additional 200 images from 200 patients were collected and used for prospective validation. For the five-category classification, the weighted average accuracy of the Inception-Resnet-v2 model reached 84.6%. The mean area under the curve (AUC) of the model for differentiating gastric cancer and neoplasm was 0.877 and 0.927, respectively. In prospective validation, the Inception-Resnet-v2 model showed lower performance compared with the endoscopist with the best performance (five-category accuracy 76.4% vs. 87.6%; cancer 76.0% vs. 97.5%; neoplasm 73.5% vs. 96.5%; P <0.001). However, there was no statistical difference between the Inception-Resnet-v2 model and the endoscopist with the worst performance in the differentiation of gastric cancer (accuracy 76.0% vs. 82.0%) and neoplasm (AUC 0.776 vs. 0.865).

Conclusion The evaluated deep-learning models have the potential for clinical application in classifying gastric cancer or neoplasm on endoscopic white-light images.

Introduction

Gastric cancer remains a global health burden and is the fourth most common cause of cancer-related death worldwide [1]. Most early gastric cancers (EGCs) lack clinical signs or symptoms and are difficult to detect and treat in a timely manner without screening strategies. Patients with premalignant lesions, such as a gastric dysplasia, also have a considerable risk of developing gastric cancer [2]. Korea has the highest incidence of gastric cancer and adopted the National Cancer Screening Program in 1999 [3]. With the widespread imple-

mentation of endoscopic screening programs, the proportion of patients with EGC at the time of diagnosis has increased [3, 4]. Although endoscopic screening programs have reduced gastric cancer mortality rates by 47 % [3], the detection of gastric neoplasms remains a challenge because it is dependent on the endoscopists' experience, expertise, and skill [5]. Moreover, repeated endoscopic examinations have been associated with decreased mortality rates from gastric cancer [3] and longer inspection times have been associated with higher proportions of neoplasm detection [6] in Korean studies, indicating that one-time screenings are not a perfect method.

Endoscopy is used for both screening and diagnosing a variety of gastrointestinal diseases, including gastric neoplasms [5]. A high quality endoscopic examination is necessary to detect malignant and premalignant lesions, especially in areas where gastric cancer is prevalent. Detection of abnormal lesions is usually based on abnormal morphology or color changes in the mucosa, and diagnostic accuracy is known to improve through training and the use of optical techniques or chromoendoscopy [5, 7, 8]. The application of endoscopic imaging technologies such as narrow-band imaging, confocal imaging or magnifying techniques (so-called image-enhanced endoscopy) is also known to enhance diagnostic accuracy [7, 9]. However, examination solely with white-light endoscopy remains the most routine form of screening, and standardization of the procedure and improvements in the interpretation process to resolve the interobserver and intraobserver variability are needed in image-enhanced endoscopy. Therefore, meticulous inspection of the stomach, discrimination of the lesions, and a targeted biopsy are the key factors in diagnosing pathologic lesions [5].

Recently, the use of convolutional neural networks (CNNs) to recognize and differentiate medical images has been increasingly adopted in clinical practice. This technique has already shown promising diagnostic performance using endoscopic images, such as detecting gastric cancer [10], recognizing *Helicobacter pylori* infection [11], classifying colorectal polyps into neoplastic or non-neoplastic features [12, 13], and distinguishing Barrett's esophagus and neoplasia [14]. However, there have been no studies on the application of CNNs in the classification of gastric neoplasms based on white-light images. This study aimed to develop a deep-learning model to automatically classify gastric neoplasms based on white-light images and to evaluate the model performance.

Methods

Study sample

All still-cut white-light endoscopy photographs of pathologically confirmed gastric lesions were retrospectively collected from consecutive patients who underwent an upper endoscopy between 2010 and 2017 at two hospitals (Chuncheon and Dongtan Sacred Heart Hospitals). The images were retrieved in JPEG format from the picture archiving and communication database system of the participant hospitals. Images were from a 35-degree field of view with a resolution of 1280 × 640 pixels. Inappropriate images were excluded according to the following exclusion criteria: 1) images with poor quality or low resolution that precluded proper classification (out of focus, artifacts, shadowing, etc.); 2) images from image-enhanced endoscopy; and 3) images without pathology results. After applying the exclusion criteria, the remaining images were included in the study. All images were de-identified by removing individual identifiers. Finally, a total of 5017 white-light images from 1269 individuals were included in the study. Of these, 812 images from 212 subjects were used as the test dataset. **Table 1s** (see the online-only supplementary material) shows the image category composition of the datasets used in the study.

The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of Chuncheon Sacred Heart Hospital (2018 – 8).

Endoscopic procedure

Upper endoscopic examinations were performed either as part of routine check-ups, for diagnosis of symptomatic patients or as a therapeutic procedure for neoplastic lesions. All of the patients fasted for more than 8 hours before the examination. All of the procedures were performed by six experienced endoscopists (> 6000 cases). Endoscopic examination was performed using the GIF-Q260, H260 or H290 endoscopes (Olympus Optical Co., Ltd., Tokyo, Japan) with an endoscopic video imaging system (Evis Lucera CV-260 SL or Elite CV-290; Olympus Optical Co.). All neoplasm-suspected lesions detected during examination were endoscopically biopsied or resected using either endoscopic mucosal resection or endoscopic submucosal dissection. Lesions that could not be resected by endoscopic procedures were surgically resected, and the final pathology results were identified. Pathological diagnosis of endoscopic biopsy was made by two specialist pathologists, each with more than 10 years of experience. Positive findings in the biopsy specimen were cross-checked with another pathologist in Chuncheon Sacred Heart hospital. The final classification of EGC, advanced gastric cancer (AGC), high grade dysplasia (HGD), or low grade dysplasia (LGD) was made by combination of histological diagnosis and clinical findings by staff members of the gastroenterology department.

Building training and test datasets

All images were reviewed by two expert endoscopists (C.S.B and S.W.P) and grouped into five categories: AGC, EGC, HGD, LGD, and non-neoplasm. The non-neoplasm category included any form of gastritis, benign ulcers, erosions, polyps, or intestinal metaplasia, etc. In addition, the images were also classified into two categories from two perspectives: cancer vs. non-cancer, and neoplasm vs. non-neoplasm. The cancer category included AGC and EGC, and the non-cancer category included HGD, LGD, and non-neoplasms. The neoplasm category included AGC, EGC, HGD, and LGD. Some images were taken of the same lesion from a different angle, direction, and distance.

The entire dataset was divided into training and test datasets, which were mutually exclusive, using random sampling. Randomization was performed based on patients and not based on images. The ratio of the patient number for training and test datasets was set to be 5:1 for each category of gastric lesions.

▶ **Table 1** Clinical characteristics of enrolled patients in the prospective validation dataset.

	No. of patients, n (%)			Age, mean (SD), years			Sex, M/F, n (% men)		
	Overall	Kangdong Sacred Heart hospital	Hallym University Sacred Heart hospital	Overall	Kangdong Sacred Heart hospital	Hallym University Sacred Heart hospital	Overall	Kangdong Sacred Heart hospital	Hallym University Sacred Heart hospital
Overall	200	88 (44.0)	112 (56.0)	62.5 (13.8)	61.3 (13.8)	61.3 (13.8)	146/54 (73.0)	68/20 (77.3)	78/34 (69.6)
AGC	28	15 (17.0)	13 (11.6)	73.8 (9.5)	69 (13.8)	79.2 (13.7)	23/5 (82.1)	14/1 (93.3)	9/4 (69.2)
EGC	46	16 (18.2)	30 (26.8)	70.5 (6.5)	72.4 (6.2)	69.4 (6.5)	34/12 (73.9)	9/7 (56.3)	25/5 (83.3)
HGD	26	14 (15.9)	12 (10.7)	63.8 (7.7)	64.6 (8.0)	62.9 (7.6)	17/9 (65.4)	10/4 (71.4)	7/5 (58.3)
LGD	30	8 (9.1)	22 (19.6)	64.2 (10.1)	65.9 (8.0)	63.6 (10.9)	25/5 (83.3)	6/2 (75.0)	19/3 (86.4)
Non-neoplasm	70	35 (39.8)	35 (31.3)	51.4 (14.1)	50.5 (13.4)	52.3 (14.9)	47/23 (67.1)	29/6 (82.9)	18/17 (51.4)

AGC, advanced gastric cancer; EGC, early gastric cancer; HGD, high grade dysplasia; LGD, low grade dysplasia; % is proportion of each category in the overall dataset.

Thus, lesions of the same category in a single patient were assigned together in one group into either training or test dataset, respectively. Of note, if a patient had lesions of different categories concurrently, the lesions could belong to different datasets because lesions of different categories were randomized independently.

The training dataset was used to fine-tune pretrained CNN models to classify gastric lesions. The test dataset, which was not balanced to ensure a similar number of lesions in each category, was subsequently used to evaluate the performance of the CNN models. The best-performing model was validated during the next stage of the study and compared with endoscopists' performance.

Prospective validation dataset

Another unused dataset was collected from two different hospitals (Kangdong- and Hallym University Sacred Heart Hospitals) to validate the established models and to compare its performance with that of three endoscopists. All still-cut white-light images of pathologically confirmed gastric lesions were prospectively collected from consecutive patients who underwent an upper endoscopy between December 2018 and February 2019 with the same exclusion criteria as those stated above. Finally, 200 images from 200 patients were selected for the prospective validation to compare the classifying performance of the established model with that of endoscopists.

After constructing the prospective validation dataset, three endoscopists (C.S.B., Y.J.Y., and J.T.H.) classified this dataset without knowing the final diagnosis. The mean experience of the endoscopists was 6.7 years (standard deviation [SD] 0.6). ▶ **Table 1** shows the characteristics of the enrolled population in the prospective validation dataset.

Constructing CNN models

Three different CNN models were used in classifying endoscopic images, namely, Inception-v4, Resnet-152, and Inception-Resnet-v2. For all CNN models, pretrained models with the ImageNet Dataset were adopted using transfer learning. Inception-v4 (https://arxiv.org/abs/1512.00567) is a revised version of CNN that achieved 21.2% top-1 and 5.6% top-5 error rates for single-frame evaluation on the ImageNet 2012 Challenge data set, which was developed and released by Google, Inc. Resnet-152 (https://arxiv.org/abs/1603.05027) is an improved version of the deep residual network, which won the ImageNet Large Scale Visual Recognition Challenge (ILSRVC) 2015 competition and surpassed human performance on the ImageNet dataset. Inception Resnet-V2 (https://arxiv.org/abs/1602.07261) is a variation of the Inception-v3 model that borrows some ideas from Microsoft's ResNet. The TensorFlow framework was adopted to implement all CNN models.

For all CNN models, a stochastic approximation of gradient descent optimization was done with the Adam optimizer. The initial learning rate, end learning rate, weight decay, and batch size were 0.01, 0.0001, $5e^{-05}$, and 30, respectively. The training dataset was preprocessed to enhance recognition performance by random cropping, resizing, flipping, and color adjustments implemented internally in each CNN model. A 5-fold cross-validation was carried out for all models, which means that the training set was further subdivided with a validation set for the selection of hyperparameters for each network. The hardware used for this study was NVIDIA's GeForce GTX 1080ti.

Main outcome measures

After constructing the CNN models using the training dataset, the performance of the models was evaluated using the test dataset and the prospective validation dataset. The main outcome

▶ **Fig. 1** Confusion matrix for per-category sensitivity of the Inception-Resnet-v2 model. **a** The test dataset. **b** The prospective validation dataset. AGC, advanced gastric cancer; EGC, early gastric cancer; HGD, high grade dysplasia; LGD, low grade dysplasia; NON, non-neoplasm.

measurements were classifying performance of established models for the five categories, gastric cancer vs. non-cancer, and gastric neoplasm vs. non-neoplasm.

Statistical methods

To investigate the performance of established CNN models, the area under the curve (AUC), was calculated. Further sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy were estimated including true positive (TP), false positive (FP), true negative (TN), and false negative (FN) values. The following formulae were used to calculate performance parameters: sensitivity TP/(TP + FN); specificity TN/(FP + TN); PPV TP/(TP + FP); NPV TN/(FN + TN); and accuracy (TP + TN)/(TP + FP + FN + TN).

Continuous variables are expressed as the mean (SD). Categorical variables are expressed as percentage with 95% confidence interval (CI). Fisher's exact test was used for the comparison of categorical variables, and DeLong test was used for the comparison of AUC values [15]. A *P*-value of < 0.05 were considered statistically significant and all tests were two-sided.

Analyses were performed using SPSS version 24.0. (IBM Corp., Armonk, New York, USA), R version 3.2.3. (R Foundation for Statistical Computing, Vienna, Austria), and Medcalc version 18.11.6 (Medcalc Software, Ostend, Belgium). The Fleiss' kappa statistic was calculated using a Microsoft Excel spreadsheet (http://www.ccitonline.org/jking/homepage/interrater.html, provided by Jason King, Ph.D.).

Results

Five-category classification performance

In the five-category classification, Inception-Resnet-v2 showed the best performance (accuracy 84.6%, 95%CI 83.69% – 85.5%) (weighted average of each class). The mean elapsed time classifying one image in the test dataset was 0.0264 seconds (SD 0.0009). The change of validation accuracy by the number of epochs is presented in **Fig. 1s**. The performance reached a plateau after the number of training epochs reached 500.

The detailed per-category performance of established models is described in **Table 2s**. The per-category AUC of established models was highest for lesions with AGC (range 0.802 – 0.855), and lowest for lesions with HGD (range 0.491 – 0.522). Of note, the per-category sensitivity was highest for lesions with non-neoplasm (range 74.2% – 92.2%); however, it was lowest for lesions with HGD (range 0 – 10.3%). The confusion matrix for the per-category sensitivity of the Inception-Resnet-v2 model in the test dataset is presented in ▶ **Fig. 1a**.

Binary classification performance

The performance of the established models in classifying gastric lesions into cancer or neoplasm is presented in **Table 3s**. In determining whether the gastric lesion was cancer or not, Inception-Resnet-v2 showed the best performance (AUC 0.877, 95%CI 0.851 – 0.901). The accuracy, sensitivity, and specificity in classifying gastric cancer were 81.9% (95%CI 79.3% – 84.6%), 75.9% (95%CI 79.3% – 84.6%), and 85.3% (95%CI 79.3% – 84.6%), respectively.

For the classification of gastric neoplasms, Inception-Resnet-v2 showed the best performance (AUC 0.927, 95%CI 0.908–0.944), and its accuracy, sensitivity, and specificity were 85.5% (95%CI 83.3%–87.8%), 84.0% (95%CI 80.5%–87.3%), and 87.3% (95%CI 83.8%–90.5%), respectively.

The AUCs for the binary classification of gastric lesions into cancer or neoplasm are presented in ▶ **Fig. 2**.

Prospective validation

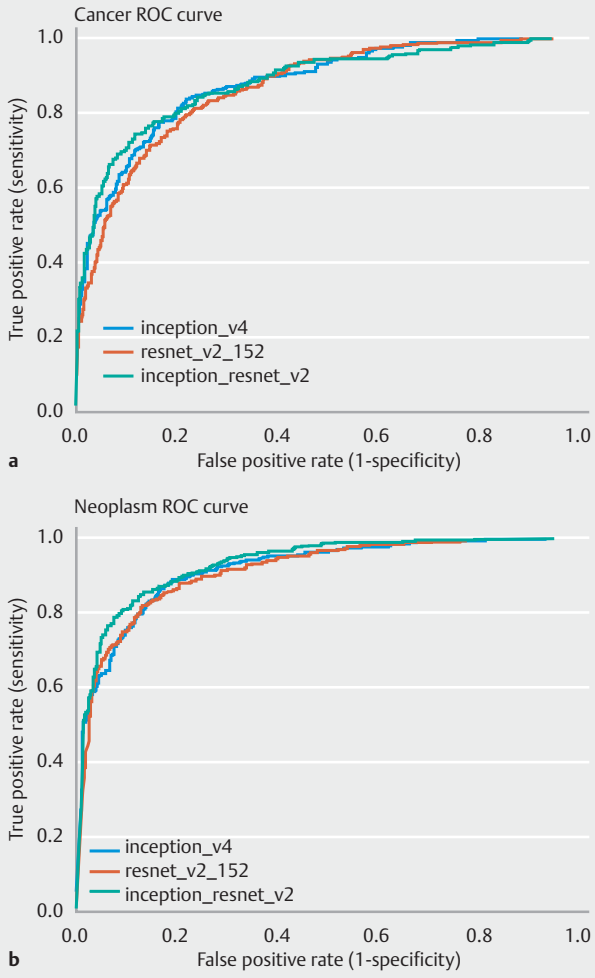
The prospective validation dataset comprised 200 images including 28 AGCs, 46 EGCs, 26 HGDs, 30 LGDs, and 70 non-neoplasms (74 cancers vs. 126 non-cancers, 130 neoplasms vs. 70 non-neoplasms). Detailed characteristics of the enrolled patients are shown in ▶ **Fig. 3** and ▶ **Table 1**.

In classifying the prospective validation dataset into five categories, an endoscopist with the best performance showed an accuracy of 87.6% (95%CI 84.3%–90.9%), whereas the Inception-Resnet-v2 model had an accuracy of only 76.4% (95%CI 72.1%–80.7%) (weighted average of each category). The performance was significantly higher for the endoscopist with the best performance than for the established model ($P<0.001$). The detailed per-category performance of endoscopists and the Inception-Resnet-v2 model is described in ▶ **Table 2**. The confusion matrix for the per-category sensitivity of Inception-Resnet-v2 model in prospective validation is described in ▶ **Fig. 1b**.

For the per-category performance of the five categories, endoscopists commonly showed the highest performance in the diagnosis of AGC (accuracy range 98.5%–99.5%) and they showed second highest diagnostic performance in the diagnosis of non-neoplasm (accuracy range 85.5%–89.5%). However, the diagnostic performance of LGD and HGD was lower than that of other lesions (accuracy range 80%–85.5%), which was common not only among endoscopists but also with the Inception-Resnet-v2 model (▶ **Table 2**).

In determining whether the gastric lesion was cancer or not, an endoscopist with the best performance showed an accuracy of 97.5% (95%CI 94.3%–99.2%), whereas the Inception-Resnet-v2 model had an accuracy of only 76.0% (9%%CI 69.5%–81.7%). The performance was significantly higher for the endoscopist with the best performance than for the established model ($P<0.001$). However, there was no statistical difference in performance in differentiating gastric cancer between the model and the two remaining endoscopists (accuracy 76.0% [95%CI 69.5%–81.7%] vs. 82.0% [95%CI 76.0%–87.1%] and 82.5% [95%CI 76.5%–87.5%]). The detailed per-category performance is described in ▶ **Table 3**.

For the classification of gastric neoplasms, the endoscopist with the best performance showed an accuracy of 96.5% (95% CI 92.9%–98.6%), whereas the Inception-Resnet-v2 model had an accuracy of only 73.5% (95%CI 66.8%–79.5%). The performance was significantly higher for the endoscopist with the best performance than for the established model ($P<0.001$). However, there was no statistical difference in differentiating neoplasm compared with the endoscopist with the worst performance (AUC 0.776 [95%CI 0.712–0.832] vs. 0.865 [95%CI 0.810–0.909]). The detailed per-category performance is described in ▶ **Table 3**.

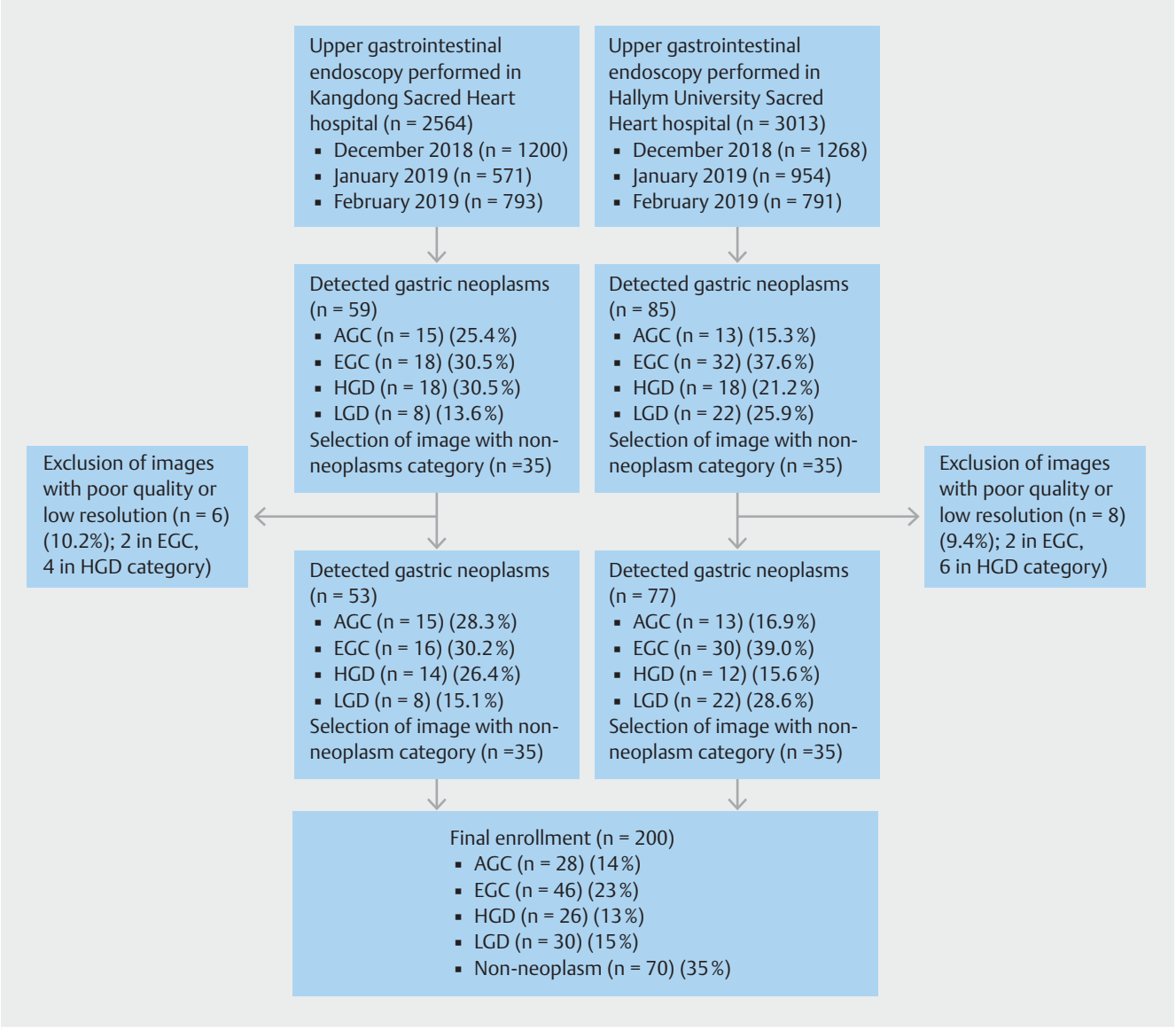


▶ **Fig. 2** Area under the curve for the prediction of: **a** gastric cancer; **b** gastric neoplasm. AUC, area under the curve; ROC, receiver operating characteristic.

Interrater reliability between three endoscopists (Fleiss's Kappa) was 0.61 for the classification of five categories ($P<0.001$), 0.64 for the classification of cancer ($P<0.001$), and 0.70 for the classification of neoplasm ($P<0.001$), which all represent substantial agreement.

Discussion

This study established the value of high performance models for the classification of gastric lesions, presenting an in situ probability of lesions, categorized into certain types, necessitating a targeted biopsy. During endoscopic screening procedures, these models may assist endoscopists in predicting the histology of ambiguous lesions and determining diagnostic or therapeutic strategies. Targeted biopsy on the presumed neoplastic lesion is the key technique for future therapeutic plans. In the case of ambiguous lesions, when it is difficult to distin-



▶ **Fig. 3** Flow diagram of prospective validation. AGC, advanced gastric cancer; EGC, early gastric cancer; HGD, high grade dysplasia; LGD, low grade dysplasia.

guish between neoplastic and non-neoplastic disease, sufficient tissue needs to be obtained for accurate diagnosis [5]. However, endoscopic biopsy is an invasive procedure that can result in mucosal damage and hemorrhage [5, 16]. Moreover, repeated biopsies or multiple biopsies over a wide area can lead to submucosal fibrosis, which impedes therapeutic procedures such as endoscopic submucosal dissection [5]. Therefore, precise prediction of the histological diagnosis during the endoscopic examination would decrease unnecessary biopsies [5].

The deep-learning models have the potential for clinical application during endoscopic procedures, although they cannot replace the procedure itself at the current time. Improving diagnostic ability during visual inspection is a constant goal for endoscopists. Although image-enhanced endoscopy has been widely adopted in clinical practice, it is not a perfect tool and more can be done to improve the accuracy of diagnosis. Automatic classification and diagnosis of ambiguous lesions can reduce unnecessary biopsies, procedures, or procedure-related adverse events [13].

Studies on the classification of gastric lesions by CNNs or other machine-learning models using endoscopic images have been limited to date. A previous study by Hirasawa et al. showed that 71 of 77 gastric cancers were correctly classified by the established CNN, with an overall sensitivity of 92.2% and a PPV of 30.6% [10]. However, the performance of this model might be overestimated because even when the CNN detected only one gastric cancer in multiple images of the same lesion, the answer was still considered to be correct [10]. The authors also claimed that all of the lesions that were missed by the CNN were superficially depressed and differentiated intramucosal cancers (corresponding to HGDs in our study) that

▶ **Table 2** Per-category diagnostic performance of three endoscopists and the established convolutional neural network model on endoscopic images in the prospective validation dataset.

Model	Diagnostic performance, % (95 %CI)					AUC (95 %CI)
	Accuracy	Sensitivity	Specificity	PPV	NPV	
Endoscopist 1						
▪ AGC	99.5 (97.3 – 99.9)	96.4 (81.7 – 99.9)	100 (97.9 – 100)	100	99.4 (96.2 – 99.9)	0.982 (0.953 – 0.996)
▪ EGC	82 (76.0 – 87.1)	56.5 (41.1 – 71.1)	89.6 (83.7 – 93.9)	61.9 (48.9 – 73.4)	87.3 (83.2 – 90.6)	0.731 (0.664 – 0.791)
▪ HGD	84 (78.2 – 88.8)	30.8 (14.3 – 51.8)	92.0 (86.9 – 95.5)	36.4 (21.0 – 55.1)	89.9 (87.3 – 92.0)	0.614 (0.542 – 0.681)
▪ LGD	84 (78.2 – 88.8)	50.0 (31.3 – 68.7)	90.0 (84.5 – 94.1)	46.9 (33.2 – 61.1)	91.1 (87.7 – 93.6)	0.700 (0.631 – 0.763)
▪ Non-neoplasm	89.5 (84.4 – 83.4)	90.0 (80.5 – 95.9)	89.2 (82.6 – 94.0)	81.8 (73.2 – 88.1)	94.3 (89.1 – 97.1)	0.896 (0.845 – 0.935)
Endoscopist 2						
▪ AGC	99 (96.4 – 99.9)	100 (87.7 – 100)	98.8 (95.9 – 99.9)	93.3 (77.9 – 98.2)	100	0.994 (0.971 – 1.000)
▪ EGC	81 (74,9 – 86.2)	47.8 (32.9 – 63.1)	90.9 (85.2 – 94.9)	61.1 (46.7 – 73.8)	85.4 (81.5 – 88.5)	0.694 (0.625 – 0.757)
▪ HGD	81.5 (75.4 – 86.6)	30.8 (14.3 – 51.8)	84.2 (78.2 – 89.2)	21.6 (12.4 – 34.9)	89.6 (86.9 – 91.8)	0.575 (0.505 – 0.643)
▪ LGD	82.5 (76.5 – 87.5)	40 (22.7 – 59.4)	90 (84.5 – 94.1)	41.4 (27.4 – 57.0)	89.5 (86.3 – 92.0)	0.650 (0.580 – 0.716)
▪ Non-neoplasm	89 (83.8 – 93.0)	90 (80.5 – 95.9)	88.5 (81.7 – 93.4)	80.8 (72.2 – 87.2)	94.3 (89.0 – 97.1)	0.892 (0.841 – 0.932)
Endoscopist 3						
▪ AGC	98.5 (95.7 – 99.7)	92.9 (76.5 – 99.1)	99.4 (96.8 – 99.9)	96.3 (78.6 – 99.5)	98.8 (95.7 – 99.7)	0.961 (0.924 – 0.983)
▪ EGC	81.5 (75.4 – 86.6)	50 (34.9 – 65.1)	90.9 (85.2 – 94.9)	62.2 (48.0 – 74.5)	85.9 (81.9 – 89.1)	0.705 (0.636 – 0.767)
▪ HGD	85.5 (79.8 – 90.1)	7.7 (0.9 – 25.1)	97.1 (93.4 – 99.1)	28.6 (7.6 – 66.2)	87.6 (86.3 – 88.8)	0.524 (0.452 – 0.595)
▪ LGD	80 (73.8 – 85.3)	56.7 (37.4 – 74.5)	84.1 (77.7 – 89.3)	38.6 (28.3 – 50.1)	91.7 (87.9 – 94.3)	0.704 (0.635 – 0.766)
▪ Non-neoplasm	85.5 (79.8 – 90.1)	90 (80.5 – 95.9)	83.1 (75.5 – 89.1)	74.1 (66 – 80.9)	93.9 (88.4 – 96.9)	0.865 (0.810 – 0.909)
Inception – Resnet – v2						
▪ AGC	93.0 (88.5 – 96.1)	60.7 (40.6 – 78.5)	98.3 (95.0 – 99.6)	85.0 (64.0 – 94.8)	93.9 (90.6 – 96.1)	0.795 (0.732 – 0.849)
▪ EGC	74.5 (67.9 – 80.4)	28.3 (16.0 – 43.5)	88.3 (82.2 – 92.9)	41.9 (27.7 – 57.6)	80.5 (77.3 – 83.3)	0.583 (0.511 – 0.652)
▪ HGD	86.4 (80.9 – 90.9)	0 (0 – 13.2)	99.4 (96.8 – 99.9)	0	86.9 (86.7 – 87.0)	0.497 (0.426 – 0.569)
▪ LGD	78.5 (72.2 – 84.0)	6.7 (0.8 – 22.1)	91.2 (85.9 – 95.0)	11.8 (3.1 – 35.6)	84.7 (83.3 – 86.0)	0.489 (0.418 – 0.561)
▪ Non-neoplasm	66.5 (59.5 – 73)	95.7 (88.0 – 99.1)	50.8 (41.9 – 59.6)	51.1 (46.6 – 55.7)	95.7 (87.8 – 98.5)	0.732 (0.665 – 0.792)

CI, confidence interval; PPV, positive predictive value; NPV, negative predictive value; AUC, area under the curve; AGC, advanced gastric cancer; EGC, early gastric cancer; HGD, high grade dysplasia; LGD, low grade dysplasia.

were difficult to distinguish from gastritis, even for experienced endoscopists [10]. However, 69.4 % of the lesions that the CNN diagnosed as gastric cancer were benign, and the most common reasons for misdiagnosis were gastritis, atrophy, and intestinal metaplasia (corresponding to non-neoplasm in our study), all of which are very common in clinical practice [10]. CNN models in our study showed highest per-category sensitivity and NPV for the non-neoplasm lesions in a prospective validation, which is presumed to be unaffected by the limitation of the previous study.

In terms of per-category performance, endoscopists demonstrated near-perfect performance in the diagnosis of AGC and the diagnostic performance of non-neoplasm also reached a substantial level. However, diagnostic performance for LGD or HGD was lower than that of other lesions, which was common not only among endoscopists but also with the Inception-Resnet-v2 model. AGC is a lesion that should not be missed by endoscopists because it is associated with significant mortality. The recognition of the characteristic morphology of AGC is emphasized during endoscopy training and the results of prospective validation are supposed to reflect endoscopists' alertness for suspected AGC lesions. However, dysplasia is defined as a lesion that refers to a mucosal structure that exhibits cytological atypia. It is categorized into low or high grade depending on

▶ **Table 3** Diagnostic performance of endoscopists and the established convolutional neural network model in classifying gastric cancer or neoplasm on endoscopic images in the prospective validation dataset.

Model	Diagnostic performance, % (95 %CI)					AUC (95 %CI)
	Accuracy	Sensitivity	Specificity	PPV	NPV	
Cancer or non-cancer						
▪ Endoscopist 1	97.5 (94.3 – 99.2)	93.2 (84.9 – 97.8)	100 (97.1 – 100)	100	96.2 (91.5 – 98.3)	0.966 (0.931 – 0.987)
▪ Endoscopist 2	82.5 (76.5 – 87.5)	74.3 (62.8 – 84.8)	87.3 (80.2 – 92.6)	77.5 (68.1 – 84.7)	85.3 (79.6 – 89.6)	0.808 (0.747 – 0.860)
▪ Endoscopist 3	82 (76.0 – 87.1)	68.9 (57.1 – 79.2)	89.7 (83.0 – 94.4)	79.7 (69.6 – 87.0)	83.1 (77.7 – 87.4)	0.793 (0.730 – 0.847)
▪ Inception-Resnet-v2	76 (69.5 – 81.7)	50 (38.1 – 61.9)	91.3 (84.9 – 95.6)	77.1 (64.7 – 86.1)	75.7 (71.1 – 79.7)	0.706 (0.638 – 0.768)
Neoplasm or non-neoplasm						
▪ Endoscopist 1	96.5 (92.9 – 98.6)	94.6 (89.2 – 97.8)	100 (97.9 – 100)	100	96 (92.2 – 98.0)	0.973 (0.948 – 0.988)
▪ Endoscopist 2	87.5 (82.1 – 91.7)	88.5 (81.7 – 93.4)	85.7 (75.3 – 92.9)	92 (86.6 – 95.3)	80 (71.1 – 86.7)	0.871 (0.816 – 0.914)
▪ Endoscopist 3	85.5 (79.8 – 90.1)	83.1 (75.5 – 89.1)	90 (80.5 – 95.9)	93.9 (88.4 – 96.9)	74.1 (66.0 – 80.9)	0.865 (0.810 – 0.909)
▪ Inception-Resnet-v2	73.5 (66.8 – 79.5)	63.8 (55.0 – 72.1)	91.4 (82.3 – 96.8)	93.3 (86.4 – 96.8)	57.7 (51.7 – 63.4)	0.776 (0.712 – 0.832)

CI, confidence interval; PPV, positive predictive value; NPV, negative predictive value; AUC, area under the curve.

the degree of atypia at cellular level. Endoscopic findings appear varied but are most commonly in the form of flat or elevated lesions, which are hard to differentiate from EGC or even from non-neoplastic lesions.

Our established models displayed weakness, especially in the classification of HGD. The reason for this relatively lower performance (five-category classification) during prospective validation is presumed to be difficulty in the diagnosis of HGD. (The AUCs of established models for the classification of HGD are commonly lowest in the test dataset and prospective validation dataset.) In real clinical practice, it is nearly impossible to accurately differentiate between HGD and EGC. Therefore, endoscopic ultrasound, image-enhanced endoscopy, or even confocal endomicroscopy are employed to resolve this issue. The number of HGD images in the validation dataset was lowest compared with the other four categories and this could have also affected the accuracy data. The establishment of models predicting depth of invasion of the lesions and enrollment of more HGD cases could resolve this issue and enhance machine learning.

The strength of this study is the enrollment of endoscopic images obtained from endoscopies performed in multiple hospitals over a long-term period and the prospective validation, which were conducted to reflect real practice patterns of endoscopists in Korea. Moreover, these models attempted to reduce the false-positive rate by presenting the probability of lesions in all types of gastric neoplasms (five categories or two categories) rather than giving only one definitive diagnosis. Moreover, binary classification with cancer vs. non-cancer, or neoplasm vs. non-neoplasm would give on-site information for the accurate prediction of gastric lesions to endoscopists and would help in determining the necessity for a biopsy specimen.

Despite the strengths, there are several limitations of the study. First, the pitfalls inherent in retrospective studies make it difficult to exclude selection bias. Some of the included images taken from an older endoscopy system had low brightness/resolution compared with the recently adopted system. Second, the performance of the CNNs presented in this study might be influenced by the composition of the database (so-called spectrum bias), although the database enrolled consecutive patients. Third, pathological classification of the lesions into five categories could be different in areas outside of Korea. No generally accepted definition has been created for differentiating gastric epithelial dysplasia or cancer, especially between Japanese and Western pathologists [17]. Although a revised Vienna classification has been proposed to address the inconsistent diagnosis of gastric epithelial dysplasias, category 4 lesions (HGD and intramucosal cancer) could still be categorized in the EGC category in some countries [17]. Therefore, the diagnostic performance could be changed if coding for images is different. Fourth, binary classification performance of the established model by prospective validation was also lower than the endoscopist with the best performance. Considering the relatively lower number of neoplasms compared with non-neoplastic lesions in the training dataset, this performance could be enhanced through the enrollment of a higher number of images with more balanced data, as we did not perform a class-balancing process in this study. Fifth, we used JPEG format for the model establishment. This compression standard is typically "lossy" and contains several user-defined settings that affect the image quality. Although JPEG was the only format that could be collected in the multicenter setting owing to technical problems, this format could induce a bias in terms of image quality. We established an unused dataset of PNG files and vali-

dated the model in a prospective manner. Further studies using only TIFF or PNG file format would avoid this type of bias. Sixth, the mechanism for the classification of the lesions by CNN is not understood, although these models showed a high diagnostic performance. This could be understood through research that explores the mechanism by dividing the images into various morphological factors that can be objectified. In summary, technical novelty and measures to minimize bias could improve the performance of established models using datasets under more realistic conditions.

In conclusion, the proposed CNNs, which classified gastric cancers/neoplasms on white-light images, displayed high performance comparable to experienced endoscopists. These evaluated models have the potential for in situ add-on testing for the accurate prediction of gastric lesions.

Acknowledgments

This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) and funded by the Korean government, Ministry of Science and ICT (MSIT) (grant number NRF2017M3A9E8033207).

Competing interests

None.

References

[1] World Health Organization Fact Sheets on Cancer. Available from: http://www.who.int/mediacentre/factsheets/fs297/en Accessed: 13 August 2018

[2] de Vries AC, van Grieken NC, Looman CW et al. Gastric cancer risk in patients with premalignant gastric lesions: a nationwide cohort study in the Netherlands. Gastroenterology 2008; 134: 945 – 952

[3] Jun JK, Choi KS, Lee HY et al. Effectiveness of the Korean National Cancer Screening Program in reducing gastric cancer mortality. Gastroenterology 2017; 152: 1319 – 1328

[4] Bang CS, Baik GH, Shin IS et al. Endoscopic submucosal dissection for early gastric cancer with undifferentiated-type histology: a meta-analysis. World J Gastroenterol 2015; 21: 6032 – 6043

[5] Bang CS, Baik GH, Kim JH et al. Effect of training in upper endoscopic biopsy. Korean J Helicobacter Up Gastrointest Res 2015; 15: 33 – 38

[6] Park JM, Huo SM, Lee HH et al. Longer observation time increases proportion of neoplasms detected by esophagogastroduodenoscopy. Gastroenterology 2017; 153: 460 – 469

[7] Muguruma N, Miyamoto H, Okahisa T et al. Endoscopic molecular imaging: status and future perspective. Clin Endosc 2013; 46: 603

[8] Cotton PB, Barkun A, Ginsberg G et al. Diagnostic endoscopy: 2020 vision. Gastrointest Endosc 2006; 64: 395 – 398

[9] Cohen J, Safdi MA, Deal SE et al. Quality indicators for esophagogastroduodenoscopy. Am J Gastroenterol 2006; 101: 886 – 891

[10] Hirasawa T, Aoyama K, Tanimoto T et al. Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images. Gastric Cancer 2018; 21: 653 – 660

[11] Itoh T, Kawahira H, Nakashima H et al. Deep learning analyzes *Helicobacter pylori* infection by upper gastrointestinal endoscopy images. Endosc Int Open 2018; 6: E139 – E144

[12] Chen PJ, Lin MC, Lai MJ et al. Accurate classification of diminutive colorectal polyps using computer-aided analysis. Gastroenterology 2018; 154: 568 – 575

[13] Komeda Y, Handa H, Watanabe T et al. Computer-aided diagnosis based on convolutional neural network system for colorectal polyp classification: preliminary experience. Oncology 2017; 93: (Suppl. 01): 30 – 34

[14] Jisu H, Bo-Yong P, Hyunjin P. Convolutional neural network classifier for distinguishing Barrett's esophagus and neoplasia endomicroscopy images. Conf Proc IEEE Eng Med Biol Soc 2017; 2017: 2892 – 2895

[15] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988; 44: 837 – 845

[16] Anderson MA, Ben-Menachem T, Gan SI et al. Management of antithrombotic agents for endoscopic procedures. Gastrointest Endosc 2009; 70: 1060 – 1070

[17] Stolte M. The new Vienna classification of epithelial neoplasia of the gastrointestinal tract: advantages and disadvantages. Virchows Arch 2003; 442: 99 – 106