Published OnlineFirst November 10, 2020; DOI: 10.1158/1078-0432.CCR-20-3159

A Prospective Validation and Observer Performance Study of a Deep Learning Algorithm for Pathologic Diagnosis of Gastric Tumors in Endoscopic Biopsies

Jeonghyuk Park[1], Bo Gun Jang[2], Yeong Won Kim[1], Hyunho Park[1], Baek-hui Kim[3], Myeung Ju Kim[4], Hyungsuk Ko[5], Jae Moon Gwak[5], Eun Ji Lee[5], Yul Ri Chung[5], Kyungdoc Kim[1], Jae Kyung Myung[6], Jeong Hwan Park[7], Dong Youl Choi[5], Chang Won Jung[5], Bong-Hee Park[5], Kyu-Hwan Jung[1], and Dong-Il Kim[5]

ABSTRACT

Purpose: Gastric cancer remains the leading cause of cancer-related deaths in Northeast Asia. Population-based endoscopic screenings in the region have yielded successful results in early detection of gastric tumors. Endoscopic screening rates are continuously increasing, and there is a need for an automatic computerized diagnostic system to reduce the diagnostic burden. In this study, we developed an algorithm to classify gastric epithelial tumors automatically and assessed its performance in a large series of gastric biopsies and its benefits as an assistance tool.

Experimental Design: Using 2,434 whole-slide images, we developed an algorithm based on convolutional neural networks to classify a gastric biopsy image into one of three categories: negative for dysplasia (NFD), tubular adenoma, or carcinoma. The performance of the algorithm was evaluated by using 7,440 biopsy specimens collected prospectively. The impact of algorithm-assisted diagnosis was assessed by six pathologists using 150 gastric biopsy cases.

Results: Diagnostic performance evaluated by the AUROC curve in the prospective study was 0.9790 for two-tier classification: negative (NFD) versus positive (all cases except NFD). When limited to epithelial tumors, the sensitivity and specificity were 1.000 and 0.9749. Algorithm-assisted digital image viewer (DV) resulted in 47% reduction in review time per image compared with DV only and 58% decrease to microscopy.

Conclusions: Our algorithm has demonstrated high accuracy in classifying epithelial tumors and its benefits as an assistance tool, which can serve as a potential screening aid system in diagnosing gastric biopsy specimens.

Introduction

Gastric cancer is the third leading cause of cancer-related deaths in both men and women worldwide (1). Although its incidence is decreasing globally, the incidence and mortality of gastric cancer remain considerably high in Northeast Asian countries, including China, Japan, and Korea (2). Gastric cancer screening is done on a population basis in Japan and Korea, and such mass screening has been shown to be effective in detecting gastric cancer at an early stage, thereby reducing mortality rates (3, 4). As a result, endoscopic screening rates for gastric cancer underwent an annual increase of 4.2% from 2004 to 2013 in Korea, reaching as high as 73.6% in the population over 40 years old in the country (5). Accordingly, the diagnostic workload for gastric biopsy specimens has increased steadily. Therefore, there is a need for an automatic computerized diagnostic system to reduce the increasing diagnostic burden and prevent misdiagnosis.

Developing an automated screening method can reduce heavy diagnostic workloads, an excellent example of which is the automated image analysis of cervical cytology specimens for cervical cancer screening (6). With advances in digital scanning devices and deep learning technologies, automated cancer diagnostic systems are being developed using whole-slide images (WSI); however, these have mostly been developed for breast and colorectal cancers (7–9). As for gastric cancers, a small number of groups have reported their automated histologic classification systems using convolutional neural networks (CNN; refs. 10–15). Sharma and colleagues trained CNNs to detect gastric cancer yielding overall classification accuracy of 69.9%; however, their dataset was limited to only 15 WSIs (12). Li and colleagues proposed a deep learning–based framework, GastricNet, for automated gastric cancer detection and demonstrated its diagnostic accuracy of 100%, a performance superior to the already well-known, state-of-the-art networks, including DenseNet and ResNet (16). However, this study was conducted only on a publicly available gastric slide dataset, lacking validation using an independent set of samples. Furthermore, gastric adenomas were not included in their study design. Yoshida and colleagues developed an image analysis software named "e-Pathologist" that classifies gastric biopsy images into either carcinoma, adenoma, or no malignancy, and performed a prospective study in a large set of gastric biopsy specimens to verify its utility (11). Although e-Pathologist could accurately identify 90.6% of negative specimens, the overall concordance rate was only 55.6%, and the

[1]VUNO Inc., Seocho-gu, Seoul, South Korea. [2]Department of Pathology, Jeju National University School of Medicine and Jeju National University Hospital, Jeju, South Korea. [3]Department of Pathology, Korea University Guro Hospital, Guro-gu, Seoul, South Korea. [4]Department of Anatomy, Dankook University College of Medicine, Chonan, Chungnam, South Korea. [5]Department of Pathology, Green Cross Laboratories, Yongin, Gyeonggi, South Korea. [6]Department of Pathology, College of Medicine, Hanyang University, Seongdong-gu, Seoul, South Korea. [7]Department of Pathology, SMG-SNU Boramae Medical Center, Seoul, South Korea.

Note: Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

J. Park, B.G. Jang, K.-H. Jung, and D.-I. Kim contributed equally to this article.

Corresponding Authors: Kyu-Hwan Jung, VUNO Inc., 6F, 507, Gangnam-daero, Seocho-gu, Seoul 06536, South Korea. Phone: 82-2-515-6646; Fax: 82-2-515-6647; E-mail: khwan.jung@vuno.co; and Dong-Il Kim, Green Cross Laboratories, Department of Pathology, 107, Ihyeonro 30 Beon-gil, Giheng-gu, Yongin-Si, Gyeonggi-do, South Korea. Phone: 82-31-260-0688; Fax: 82-31-260-9609; E-mail: ilkim@gclabs.co.kr

Clin Cancer Res 2021;27:719–28

doi: 10.1158/1078-0432.CCR-20-3159

Published OnlineFirst November 10, 2020; DOI: 10.1158/1078-0432.CCR-20-3159

Translational Relevance

Diagnostic workload for gastric biopsy specimens is increasing steadily in Northeast Asia. Previous studies on deep learning–assisted analysis of digital gastric biopsy images have limitations with respect to clinical validation. In this study, we developed and evaluated a deep learning algorithm for histologic classification of gastric epithelial tumors in actual clinical practice. Our algorithm demonstrated a superb performance with diagnostic accuracy of 0.97–0.99 calculated from the AUROC curve in both retrospective and prospective studies. Algorithm-assisted diagnosis significantly saved the average review time per image, particularly for negative cases, reaching 100% sensitivity. We believe that our algorithm can potentially serve as a screening or an assistance tool not only in countries with heavy diagnostic workloads of gastric biopsy specimens, but also in areas where experienced pathologists are not available.

false-negative rate was as high as 9.4%. Iizuka and colleagues evaluated their algorithm that classifies WSIs into either nonneoplastic, adenoma, or adenocarcinoma, and confirmed accuracy of 95.6% from their test set consisting of 45 WSIs (13). Although the aforementioned CNN-based gastric cancer detection algorithms have shown promising results, their diagnostic performance in the real clinical setting, as well as impact on diagnostic workflow have not been validated.

In this study, we aimed to develop and evaluate a high-performance deep learning algorithm for automated histologic classification of gastric epithelial tumors (gastric adenomas and carcinomas) for endoscopic biopsy specimens. Using a large dataset consisting of 2,434 WSIs, we trained a baseline CNN and improved its performance by applying additional color/space augmentation and representation aggregation. Furthermore, the performance of our proposed model was evaluated with 7,440 WSIs prospectively, and its clinical benefits as an assistance tool were assessed in an observer study. As a result, our algorithm demonstrated consistent outstanding accuracy and efficiency in identifying gastric epithelial tumors in both prospective and observer studies, proving its potential as a screening or assistance tool for gastric biopsy diagnosis.

Materials and Methods

Dataset

Training and validation set

To train the algorithm for gastric epithelial tumors, a total of 1,678 cases of gastric resection ($n = 201$, 12%) and endoscopic biopsy specimens ($n = 1,477$, 88%) from 1,522 patients were collected from two institutions: Korea University Guro Hospital (Guro-gu, Seoul, South Korea, KUGH), and Green Cross Laboratories (GCL). A total of 792 (52%) were male and 730 (48%) were female, with an age range of 28–89 years old (mean $\pm$ SD, 60 ± 13). For retrospective evaluation, we collected 756 cases of endoscopic biopsy specimens from GCL and Jeju National University Hospital (Jeju, South Korea, JNUH); 392 (52%) were male and 364 (48%) were female, with an age range of 27–92 years old (mean $\pm$ SD, 59 ± 14). We call this retrospective evaluation dataset as the validation set hereafter..

Test and observer study set

For the prospective study, GCL accrued 7,459 consecutive gastric biopsies from 5,393 patients who received gastroscopy at eight local clinics or hospitals located in Gyeonggi-do province in South Korea from July 2019 to November 2019. The purpose of gastroscopy was gastric cancer screening in 72% (3,883/5,393) of patients, and the remaining 28% (1,510/5,393) underwent the procedure for the diagnosis of gastrointestinal (GI) symptoms, such as heartburn and indigestion. A total of 2,754 (51%) were male and 2,639 (49%) were female, with an age range of 29–93 years old (mean $\pm$ SD, 58 ± 12). We call this prospective evaluation dataset as the test set hereafter. For the observer study, 150 cases of endoscopic biopsy specimens were collected from 150 patients from GCL; 81 (54%) were male and 69 (46%) were female, with an age range of 25–87 years old (mean $\pm$ SD, 62 ± 14). All specimens were made into glass slides of formalin-fixed, paraffin-embedded tissue stained with hematoxylin and eosin (H&E) using an automated staining system. Patient information was removed from all slides for deidentification. All slides were scanned with a virtual slide scanner Aperio VERSA (Aperio) at 40 $\times$ magnification. This study was approved by the Institutional Review Board (IRB) at the KUGH (Guro-gu, Seoul, South Korea, IRB no., 2017-GR0792), GCL (IRB no., GCL-2017-2002-06), and JNUH (Jeju, South Korea, IRB no., 2017-11-014), respectively, and it was conducted in accordance with the Declaration of Helsinki. Informed consent from the patients was waived with IRB approval.

Pathologic diagnosis

Three experienced GI pathologists at GCL (D.-I. Kim, J.M. Gwak, and H. Ko) evaluated each slide independently and reached a consensus for the reference diagnosis. We trained our algorithm to classify each region into three categories: negative for dysplasia (NFD), tubular adenoma (TA), and carcinoma (CA). Comparison of this categorization with Korean pathologists' diagnosis and the revised Vienna classification (17) is summarized in Supplementary Table S1. As for the revised Vienna classification, category 1 is NFD. Category 3 and categories 4.1–4.2 belong to TA, and categories 4.3–5.2 are CA. The algorithm's classification does not cover category 2 because it is only used when one cannot decide whether a lesion is nonneoplastic or neoplastic (17). The training set included 1,678 cases (1,218 NFD, 187 TA, and 273 CA cases) and the validation set included 756 cases (428 NFD, 162 TA, and 166 CA cases), as shown in Supplementary Table S2. For the test set, 7,440 cases were included after excluding 19 cases with too small biopsy size. In total, there were 6,441 cases of chronic gastritis, 838 cases of nonneoplastic polyp (574 fundic gland polyp [FGP], 251 hyperplastic polyp (HP), 11 xanthoma, one inflammatory fibroid polyp, and one heterotopic pancreas), 81 cases of TA, and 64 cases of CA (**Table 1**). In addition, eight cases were diagnosed as indefinite for dysplasia (IFD), and eight cases as nonepithelial tumors, including mucosa-associated lymphoid tissue (MALT) lymphoma ($n = 5$), neuroendocrine tumor ($n = 2$), and GI stromal tumor (GIST, $n = 1$). For all 145 gastric epithelial tumors, there was no disagreement in diagnosis among three pathologists. Our observer study of 150 cases (120 NFD, 15 TA, and 15 CA cases) was divided into three sets: set 1 included 40 NFD, five TA, and five CA, set 2 included 40 NFD, four TA, and six CA, and set 3 included 40 NFD, six TA, and four CA cases, respectively. A summary of the sets is presented in Supplementary Table S3. For each slide, both region-level and slide-level annotations were made. The region-level annotation was made by an experienced GI pathologist by marking the regions of interest that were pathognomonic (Supplementary Fig. S1). Slide-level annotations were made with the consensus of at least two pathologists. Brief information on the study workflow is visualized in **Fig. 1A**.

Published OnlineFirst November 10, 2020; DOI: 10.1158/1078-0432.CCR-20-3159

Table 1. Diagnostic classification by human pathologists and proposed algorithm.

	Classification by algorithm							
	Validation set (n = 756)			Total	Test set (n = 7,440)			
Final diagnosis by pathologists	NFD	TA	CA		NFD	TA	CA	Total
CG [no. (%)]	421 (98.4)	5 (1.1)	2 (0.5)	428 (56.6)	6,298 (97.8)	80 (1.2)	63 (1.0)	6,441 (86.4)
Nonneoplastic polyp [no. (%)]	—	—	—					
Fundic gland polyp	—	—	—		547 (95.3)	24 (4.2)	3 (0.5)	574 (7.7)
Hyperplastic polyp	—	—	—		239 (95.2)	11 (4.4)	1 (0.4)	251 (3.4)
Xanthoma	—	—	—		10 (90.9)	0 (0.0)	1 (1.1)	11 (0.1)
Inflammatory fibroid polyp	—	—	—		1 (100.0)	0 (0.0)	0 (0.0)	1 (<0.1)
Heterotopic pancreas	—	—	—		1 (100.0)	0 (0.0)	0 (0.0)	1 (<0.1)
TA [no. (%)]								
TA, LGD	2 (1.3)	145 (97.3)	2 (1.3)	149 (19.7)	0 (0.0)	65 (97.0)	2 (3.0)	67 (0.9)
TA, HGD	0 (0.0)	9 (69.2)	4 (30.8)	13 (1.7)	0 (0.0)	10 (71.4)	4 (28.6)	14 (0.2)
CA [no. (%)]	1 (0.6)	1 (0.6)	164 (98.8)	166 (22.0)	0 (0.0)	1 (1.6)	63 (98.4)	64 (0.9)
Indefinite for dysplasia [no. (%)]					4 (50.0)	2 (25.0)	2 (25.0)	8 (0.1)
Others [no. (%)]								
MALT lymphoma	—	—	—		4 (80.0)	0 (0.0)	1 (20.0)	5 (<0.1)
Neuroendocrine tumor	—	—	—		0 (0.0)	0 (0.0)	2 (100.0)	2 (<0.1)
GIST	—	—	—		1 (100.0)	0 (0.0)	0 (0.0)	1 (<0.1)
Total [no. (%)]	424 (56.1)	160 (21.1)	172 (22.8)	756 (100.0)	7,105(95.5)	193(2.6)	142(1.9)	7,440 (100.0)

Abbreviations: CG, chronic gastritis; HGD, high-grade dysplasia; LGD, low-grade dysplasia.

Observer study

A multicenter, reader-blinded study was performed with participation from six pathologists at four different institutions in South Korea (GCL; JNUH, Jeju, South Korea; Hanyang University Hospital, Seongdong-gu, Seoul, South Korea; and Boramae Hospital, Sindaebang-dong, Seoul, South Korea). The six pathologists have a minimum of 5 years of surgical pathology experience and have never used digital pathology in clinical practice before. They did not participate in case collection or establishing reference diagnosis. They were instructed to review the slides and images at a rate similar to their routine practice. For the observer study dataset, 150 gastric biopsy specimens were obtained from GCL, which were tumor enriched with a tumor prevalence of 20% compared with test set.

A total of 150 cases were further divided into three sets (each set contained 20% of positive cases) and were evaluated by the pathologists under an inspector's supervision who managed and recorded the process. Each pathologist was randomly designated to assess the three sets by conventional microscope (Mic), digital image viewer (DV), and algorithm-assisted DV (AADV). To reduce and evaluate the effect of possible bias from the dataset and reading method, the observer study was designed to assign different datasets and reading methods for each observer to cover all possible combinations. To establish familiarity with the DV (with or without algorithm), a review of five training images that were not part of the study cases was conducted before each session. During sessions with DV and AADV, the time from opening the image in the viewer to opening the next image was measured. During sessions with Mic, the time from placing the slide on the microscope's stage to placing the next one was measured from the video recording.

Proposed framework

In accordance with the redefined three categories, NFD, TA, CA, we applied our proposed framework to generate both region-level and slide-level classification from the WSIs of H&E-stained gastric biopsy specimens (**Fig. 1B** and **C**). As the first step, the baseline CNN was trained to extract spatially reduced features or representations using the image and annotations of small patches from the WSI at 10× scale (see Baseline Framework section in Supplementary Data for details). After the feature extraction phase, additional convolution layers were applied to expand the receptive field of the model to utilize wider high-level context for more accurate region-level predictions. We called this modified model a representation aggregation CNN (RACNN). To generate a slide-level prediction, the summation of region-level predictions was calculated and used as a slide-level feature. Using these sets of slide-level features and their corresponding slide-level annotations, a random forest classifier was trained to classify each WSI into a slide-level category. To further improve the performance and robustness of the proposed pipeline, we adopted stain normalization and CIELAB color space augmentation in training feature extraction model of our RACNN model (see Stain Normalization and CIELAB Color Space Augmentation sections in Supplementary Data for details).

Results

Table 1 summarizes the final diagnosis by the pathologists and classification by the algorithm for the gastric biopsy specimens. Representative histologic image for each lesion and the corresponding heatmap generated by the algorithm are presented in **Fig. 2**, showing significant overlap between the tumor areas marked by the pathologist and those detected by the algorithm.

Performance of algorithm in the validation set

We used the AUROC curve, as well as sensitivity and specificity to assess the performance of the algorithm quantitatively. For 756 gastric biopsy specimens containing 328 (43%) cases of TA or CA, our algorithm demonstrated remarkably high accuracy. For a two-tier classification, that is, negative (NFD) versus positive, AUROC, sensitivity, and specificity were 0.9949 [95% confidence interval (CI), 0.9890–0.9995], 0.9909 (95% CI, 0.9808–1.0000), and 0.9813 (95% CI, 0.9682–0.9929), respectively (**Table 2**; **Fig. 3A**). For a three-tier classification (NFD vs. TA vs. CA), the macro-averaged AUROC (the mean value of AUROCs for NFD, TA, and CA) was 0.9922 (95% CI, 0.9828–0.9986). Overall accuracy and balanced accuracy are shown in **Table 2**.

Published OnlineFirst November 10, 2020; DOI: 10.1158/1078-0432.CCR-20-3159

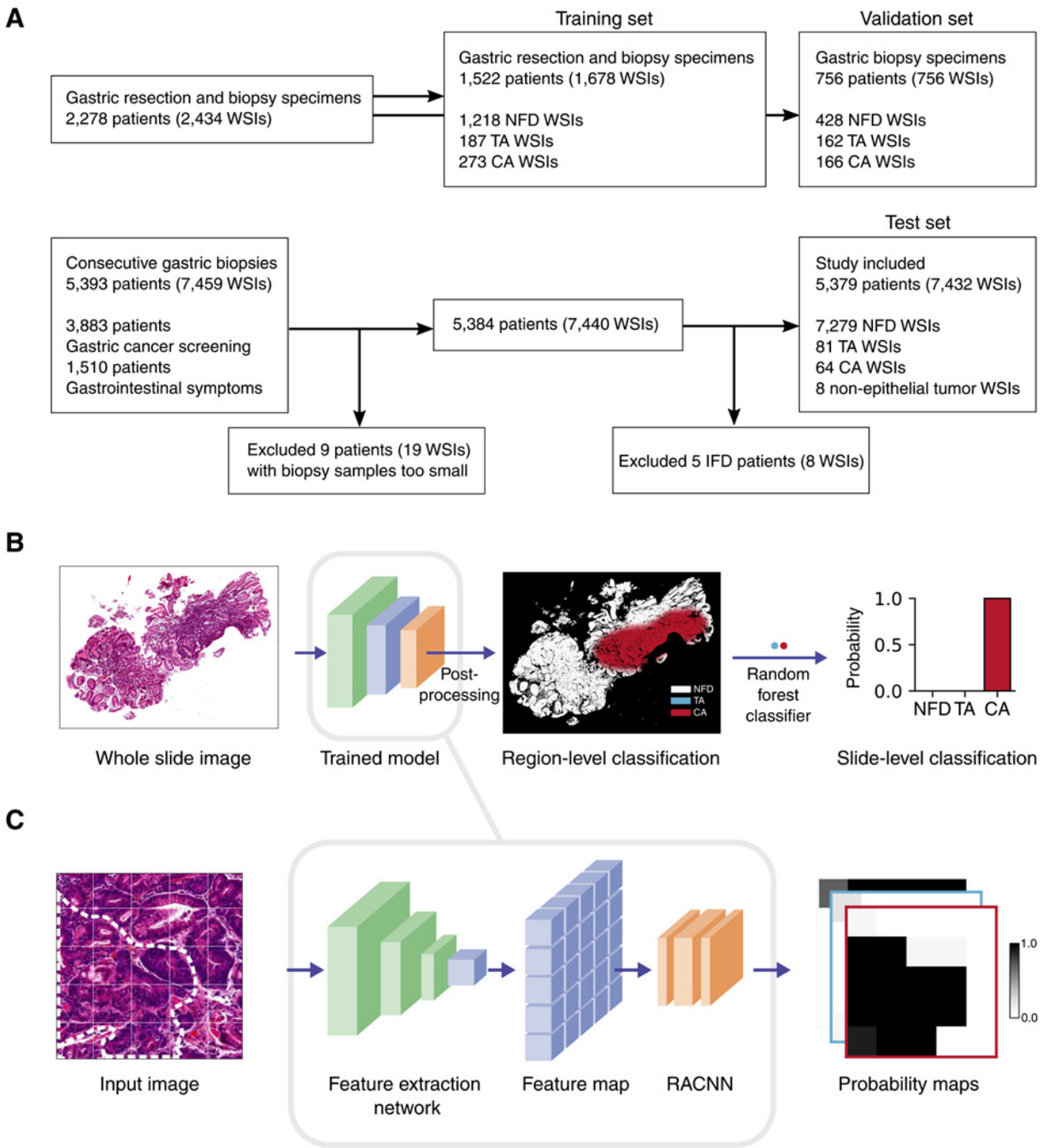

Figure 1.
Datasets and the proposed framework. **A,** Study profile of training set, validation set, and test set. **B,** The proposed framework. **C,** The architecture of the feature extraction network and RACNN.

Performance of the algorithm in the test set

To test our algorithm's performance in real practice, we conducted a 5-month long prospective study at GCL. A large number of gastric biopsy specimens ($n = 7{,}440$) was collected from 5,384 patients, and a parallel diagnostic process was carried out by our pathologists and the algorithm. For the two-tier classification, AUROC, sensitivity, and specificity were 0.9790 (95% CI, 0.9612–0.9932), 0.9673 (95% CI, 0.9365–0.9930), and 0.9749 (95% CI, 0.9713–0.9785), respectively (**Table 2**; **Fig. 3B**). For the three-tier classification, the macro-averaged AUROC was 0.9742 (95% CI, 0.9494–0.9935; **Table 2**). When the cases were confined to only epithelial tumors, the sensitivity of our algorithm reached 1.0000 (95% CI, 1.0000–1.0000; **Table 2**; **Fig. 3C**).

Published OnlineFirst November 10, 2020; DOI: 10.1158/1078-0432.CCR-20-3159

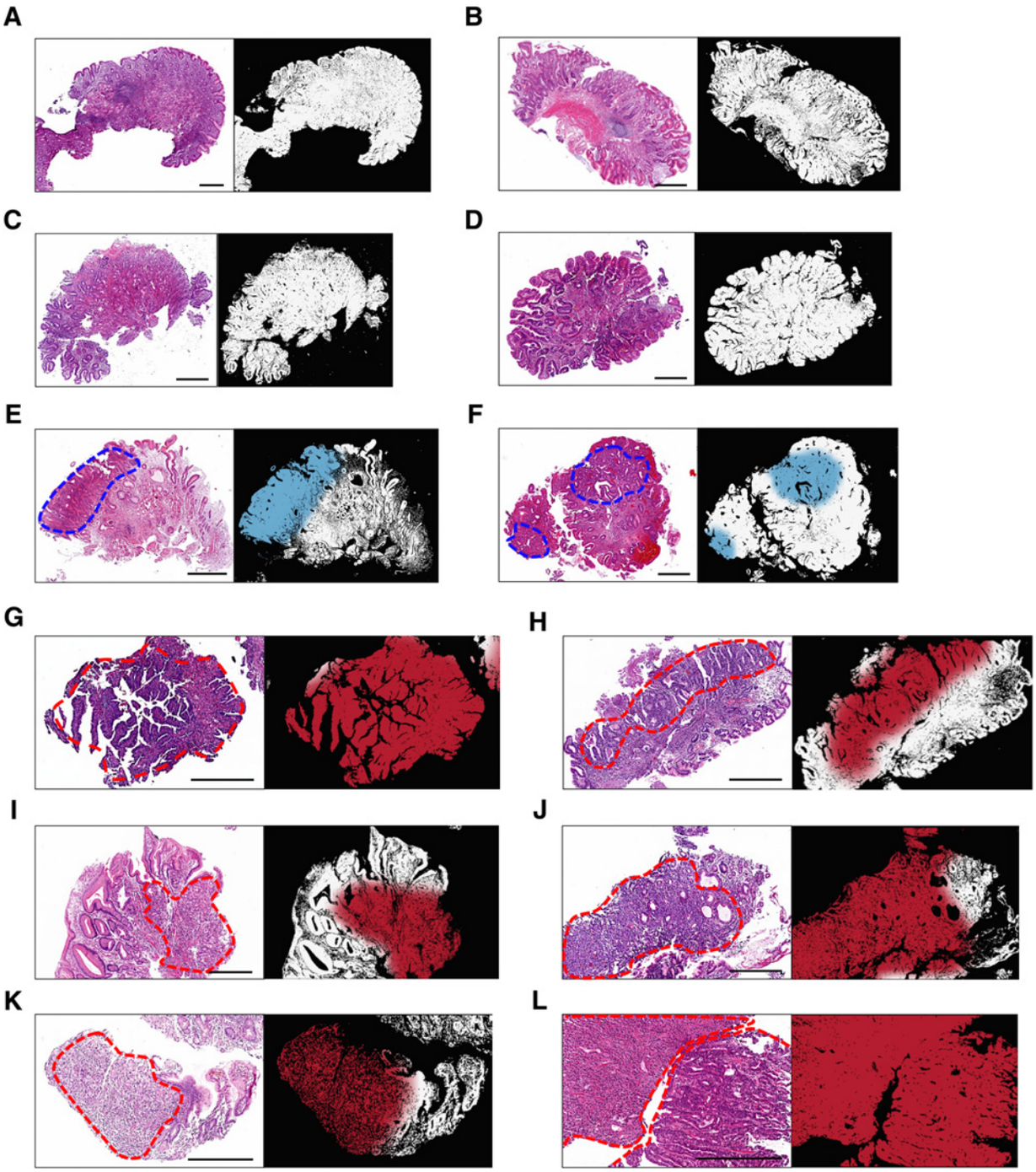

Figure 2.

Gastric biopsies interpreted as NFD include gastric mucosa from the fundus (**A**) and antrum (**B**), gastric mucosa with erosion (**C**), and intestinal metaplasia (**D**). **E** and **F,** Gastric adenomas with low-grade dysplasia, marked by the dashed blue lines, were recognized and depicted as blue areas by the algorithm. Images of papillary adenocarcinoma (**G**), tubular adenocarcinomas with well (**H**), moderate (**I**), and poor (**J**) differentiation, SRCC (**K**), and mixed carcinoma (**L**). Carcinoma areas identified by the pathologists and marked by the red dashed lines correspond well with the red areas recognized by the algorithm. Scale bar, 500 μm.

False-negative predictions by the algorithm

No false-negative prediction was made by the algorithm in the prospective test set. Three of 756 (0.4%) cases were found to be false negative in the validation set: two CAs were classified as TA and NFD, and one TA as NFD, as summarized in Supplementary Table S4. Interestingly, a signet ring cell carcinoma (SRCC) was classified as NFD despite a large area with the cancer cells in the biopsy tissue (Supplementary Fig. S2A). On reviewing the slide, it appeared that the

Published OnlineFirst November 10, 2020; DOI: 10.1158/1078-0432.CCR-20-3159

Table 2. Performance of proposed algorithm in the validation and test set.

	Validation set ($n = 756$)		Test set ($n = 7,440$)			
Metric	**Two-tier ($n = 756$)**	**Three-tier ($n = 756$)**	**Two-tier ($n = 7,432$)[a]**	**Three-tier ($n = 7,432$)[a]**	**Two-tier ($n = 7,424$)[b]**	**Three-tier ($n = 7,424$)[b]**
Sensitivity (95% CI)	0.9909 (0.9808–1.0000)		0.9673 (0.9365–0.9930)		1.0000 (1.0000–1.0000)	
Specificity (95% CI)	0.9813 (0.9682–0.9929)		0.9749 (0.9713–0.9785)		0.9749 (0.9713–0.9784)	
AUROC (95% CI)	0.9949 (0.9890–0.9995)	0.9922 (0.9828–0.9986)[c]	0.9790 (0.9612–0.9932)	0.9742 (0.9494–0.9935)[c]	0.9972 (0.9962–0.9981)	0.9928 (0.9854–0.9981)[c]
Accuracy (95% CI)	0.9854 (0.9762–0.9934)	0.9775 (0.9669–0.9868)	0.9747 (0.9711–0.9783)	0.9735 (0.9699–0.9773)	0.9754 (0.9718–0.9789)	0.9741 (0.9706–0.9779)
Balanced accuracy (95% CI)	0.9861 (0.9777–0.9937)	0.9741 (0.9612–0.9863)	0.9711 (0.9551–0.9836)	0.9309 (0.8975–0.9591)	0.9874 (0.9857–0.9892)	0.9535 (0.9267–0.9772)

Note: Two-tier classification refers to NFD vs. the rest; three-tier classification refers to NFD vs. TA vs. carcinoma.

[a]All cases except indefinite for dysplasia.

[b]All cases except indefinite for dysplasia, MALT lymphoma, neuroendocrine tumor, and GIST.

[c]Macro-averaged AUROC.

mild nuclear atypia of the cancer cells might have contributed to the misclassification. Two typical TA cases were classified as NFD, probably due to the less glandular crowding from stromal edema (Supplementary Fig. S2B).

False-positive predictions by the algorithm

Seven false-positive cases (1%) were observed in the validation set (**Table 1**; Supplementary Table S5), and 183 false-positive predictions (2.5%) were made in the training set; 115 cases of chronic gastritis or nonneoplastic polyps were classified to TA and 68 cases to CA (**Table 1**). The 115 false TA diagnoses by the algorithm included 80 chronic gastritis, 24 FGP, and 11 HP cases, and 71 false CA cases included 63 chronic gastritis, three fundic gland polyps, a hyperplastic polyp, a xanthoma, a MALT lymphoma, and two neuroendocrine tumors. The probable causes of misdiagnosis by algorithm are summarized in Supplementary Table S6. The most common cause of misprediction into TA was regenerative atypia (38.3%) followed by intestinal metaplasia (18.2%). On the other hand, inflammatory tissues (50%), such as ulcer detritus and granulation tissue, were the most common histologic features that likely led to misclassification into CA. Representative false-positive cases are shown in Supplementary Fig. S3.

Observer study

To evaluate the potential impact of our algorithm on deep learning–supported diagnosis, we conducted a multicenter, reader-blinded study for classification of gastric biopsies. Three sets of gastric biopsy cases ($n = 50$/set) were independently classified by six pathologists by using three different modalities, as shown in **Fig. 4A**; the time from the start of the review to establishing a diagnosis was recorded. There was no significant difference in overall review time per slide for the three sets (Supplementary Fig. S4A). The summary of results is shown in Supplementary Table S7. For a two-tier classification, NFD versus positive, our algorithm showed 0.9600 (95% CI, 0.9600–0.9600) accuracy, 1.0000 (95% CI, 1.0000–1.0000) sensitivity, and 0.8333 (95% CI, 0.8333–0.8333) specificity. Diagnostic performance was compared among the three groups by modality: Mic group, DV group, and AADV group. Overall accuracy was 0.9633 (95% CI, 0.9226–1.0000) in Mic group, 0.9867 (95% CI, 0.9578–0.9975) in DV group, and 0.9967 (95% CI, 0.9881–1.0000) in AADV group. No significant difference in accuracy, sensitivity, or specificity was observed among the three groups (**Fig. 4B**). Algorithm-alone classification and AADV group reached sensitivity of 1.0000.

The average time of review per slide was shortest in algorithm-assisted group (**Fig. 4C**), with 44.97 (95% CI, 41.43–48.52), 35.70 (95% CI, 33.24–38.15), and 18.90 (95% CI, 17.44–20.36) seconds in Mic, DV, and AADV groups, respectively. The same results were observed when the time was normalized by mean review time of each pathologist (Supplementary Fig. S4B), and all six pathologists yielded consistent results (Supplementary Fig. S4C). In particular, the difference in review time was more apparent in the negative cases ($P < 0.001$ for Mic vs. AADV and DV vs. AADV); 46.78 (95% CI, 42.68–50.87), 36.08 (95% CI, 33.57–38.58), and 17.37 (95% CI, 15.85–18.89) seconds in Mic, DV, and AADV groups. For TA/CA cases, review time was also shorter in AADV group than Mic group ($P < 0.05$) with 37.77 (95% CI, 31.07–44.46), 34.18 (95% CI, 26.91–41.46), and 25.02 (95% CI, 21.25–28.79) seconds in Mic, DV, and AADV groups, however, there was no significant difference in review time between DV and AADV groups ($P = 0.09$). These findings demonstrate that the reduced reading time by algorithm support is mostly attributed to the time saved from reviewing negative cases.

Published OnlineFirst November 10, 2020; DOI: 10.1158/1078-0432.CCR-20-3159

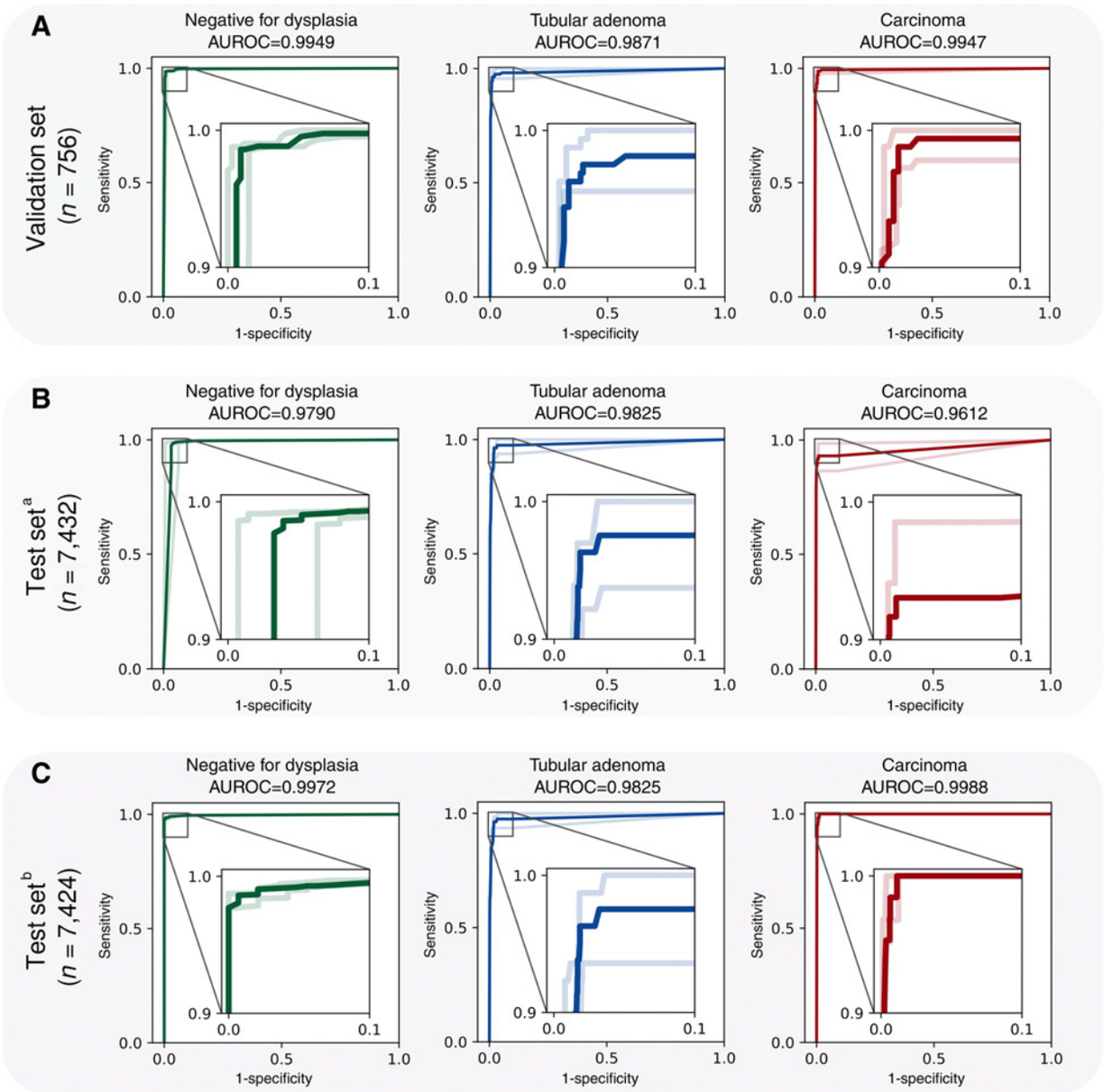

Figure 3.

ROC curves in the validation and test sets for each gastric lesion: NFD), TA, and CA in the validation set (**A**), test set[a] (**B**), and test set[b] (**C**). Note that ROC curve for NFD is equivalent to two-tier classification (NFD vs. the rest). Test set[a], all cases except indefinite for dysplasia and test set[b], all cases except indefinite for dysplasia, MALT lymphoma, neuroendocrine tumor, and GIST. The dim lines correspond to the top and bottom bounds of 95% CI.

Discussion

Recent studies have shown promising results of deep learning–based algorithms in diagnosing pathologic lesions in digitized H&E slides from various organs, including the stomach, breast, skin, prostate, and lung cancers (9, 11, 14, 18–20). As for gastric cancer, the increasing diagnostic workload of endoscopic biopsy specimens calls for development of high-performance algorithm with high sensitivity and specificity. In this study, we developed an algorithm to detect gastric tumors and demonstrated a superb performance by showing 100% sensitivity and 97% specificity in the prospective study when limited to epithelial tumors (**Table 2**). Even when nonepithelial tumors were included, it showed a 96% sensitivity, 97% specificity, and 97% accuracy. Considering that the clinically significant diagnostic error rate in surgical pathology has been reported to vary from 0.26% to 1.2%, our algorithm's accuracy in diagnosing gastric tumors seems to be almost equal to that of human pathologists.

Studies have suggested the utility of deep learning algorithms in assisting pathologists to improve accuracy and efficiency in cancer diagnosis (21, 22). To investigate the potential benefit of our algorithm as an assistance tool for interpreting gastric biopsies, we performed an observer study using three different modes: traditional microscope

Published OnlineFirst November 10, 2020; DOI: 10.1158/1078-0432.CCR-20-3159

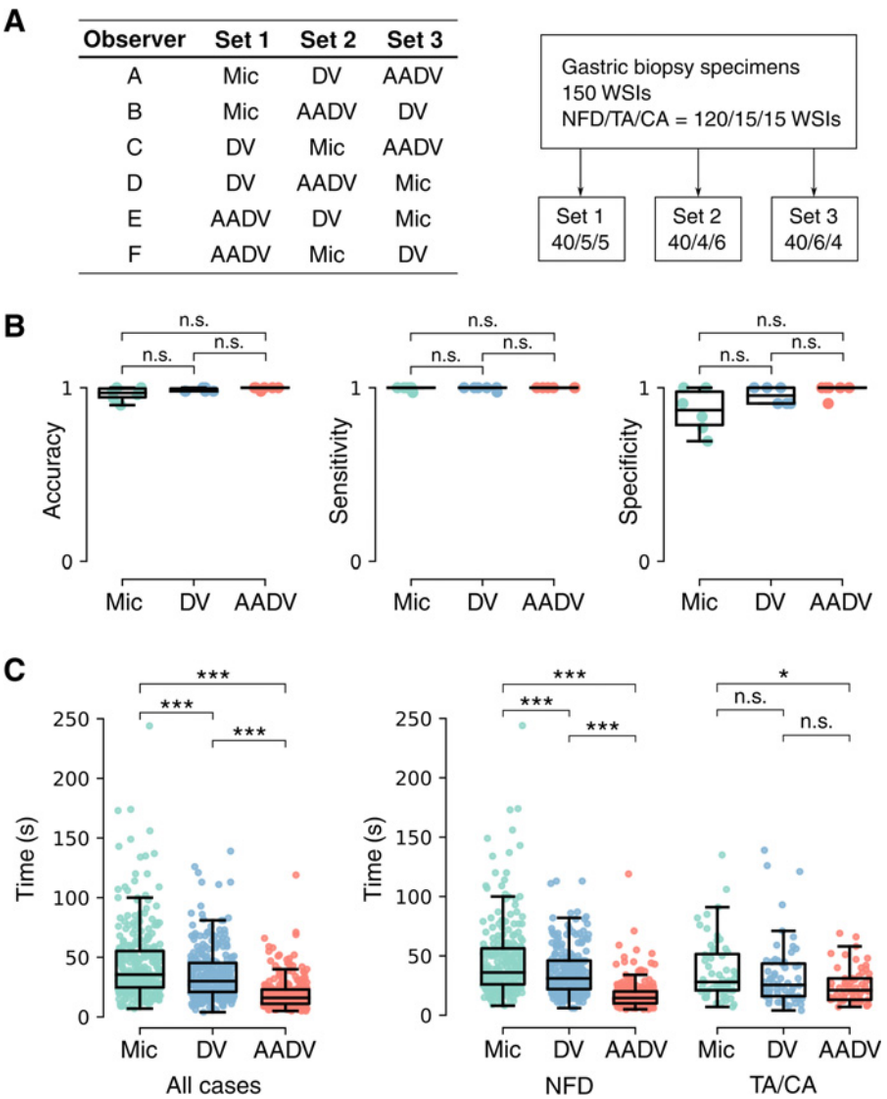

Figure 4.

Observer performance and review time per slide with assistance. **A,** Observer study design (left) and profile of study set (right). **B,** Observer performance. Black circles represent each observer. Accuracy (left), sensitivity (center), and specificity (right) are shown. **C,** The review time of all cases (left) and NFD and TA/CA (cancer) cases (right). Colored circles represent each slide. *, $P < 0.05$; ***, $P < 0.001$, ANOVA with *post hoc* Tukey HSD test. n.s., not significant.

(Mic), DV, and AADV. Although the overall accuracy was higher in AADV group (0.9967) than in Mic group (0.9633), the difference was not statistically significant ($P = 0.07$). As disappointing as it may look, we believe that this is primarily due to the high accuracy of pathologists for gastric cancers in endoscopic biopsies. According to one study including 1,331 patients who received endoscopic biopsy, overall diagnostic accuracy for gastric cancer was 97.4% (23). Our algorithm's diagnostic accuracy was in the range of 0.97–0.98 in the validation and test sets, and it even achieved 0.99 in the observer study. In addition, we found that only the AADV group reached 100% sensitivity, whereas Mic and DV groups each missed a positive case. Both of these cases had tiny cancer foci present in the biopsy specimen. This suggests that the high sensitivity achieved by algorithm assistance could reduce the risk of missing cancer, particularly in biopsies with a small area of cancer cells.

Some studies have already demonstrated that artificial intelligence (AI)-based algorithms can even exceed the sensitivity of pathologists in detecting cancer foci in digital images. However, it comes at the cost of increased false positivity (18). In our prospective study, 183 false predictions (115 false TAs and 68 false CAs) were made by the algorithm, while achieving 100% sensitivity, and algorithm-alone classification had relatively low specificity of 0.8333 in the observer study. Notably, however, in the observer study, all false-positive cases, except one case, were corrected by the pathologists during the review process, and the AADV group resulted in the highest specificity. Upon reviewing the false-positive cases, we found that inflammation- or ulcer-induced cellular atypia was responsible for most of the false classifications by the algorithm (Supplementary Table S6). Although reactive atypia often demonstrates histologic features almost identical (or sometimes even worse) to those observed in gastric cancer cells, experienced pathologists can easily identify them as benign from the surrounding histologic context. Furthermore, in a screening or an assisted mode, it is more important for an algorithm not to miss a tumor than to avoid a false-positive diagnosis. Taken together, these findings suggest that our algorithm as an assisting tool has the potential to maximize both sensitivity and specificity in detecting tumors in gastric biopsy specimens.

Published OnlineFirst November 10, 2020; DOI: 10.1158/1078-0432.CCR-20-3159

There are several deep learning models that have been developed for the diagnosis of gastric cancers using WSIs (11–13, 15, 24), and a comparison of some of these studies and ours is summarized in Supplementary Table S9. Although each study adopted a different type of algorithm, most of them demonstrated a powerful diagnostic accuracy. For example, Iizuka and colleagues (13) have reported 0.97–0.98 AUROC, which is comparable with our results of 0.97–0.99 AUROC in the prospective set. Yoshida and colleagues were the first to test AI algorithm for gastric cancer diagnosis in the biopsy specimens, but its accuracy was not satisfactory (11). Most recently, Song and colleagues have shown a high sensitivity and accuracy of AI assistance system by conducting a multicenter test (15). However, this study included both surgical and biopsy specimens and only junior pathologists participated in the observer study to show that AI assistance helped the pathologists achieve better accuracy. On the other hand, in our study, we only included biopsy specimens in all sets, except training set, and involved experienced pathologists to avoid overestimating the algorithm's performance compared with human pathologists.

More importantly, we provided the evidence of time-saving benefits of algorithm when it is applied as an assistant tool, which is one of the critical factors that must be tested to determine the applicability of an algorithm in real diagnostic workflow. Overall, algorithm assistance resulted in a 47% reduction in review time per image compared with DV group and a 58% reduction compared with Mic group. The time-saving effect was more apparent in NFD cases (52% reduction; $P < 0.001$). Although a 27% reduction was observed in TA/CA cases, it was not statistically significant ($P = 0.09$). This increased efficiency for diagnosing negative cases is notable given that we included 20% of TA/CA cases in observer study set. In our prospective study including 7,440 endoscopic biopsies collected over 5 months, only 2% (153 cases) of them turned out to be gastric tumors that required additional endoscopic or surgical intervention. Therefore, algorithm assistance would have saved a considerable amount of time if it had been applied to our actual clinical practice with negative cases comprising more than 98% of specimens.

Although our algorithm successfully classified all SRCCs in the prospective study, one case of SRCC in the validation set yielded a false-negative result, which was misclassified as NFD (Supplementary Fig. S2A). Because of its deceptively bland morphology, diagnosing SRCC in small biopsies can be challenging even for pathologists at times. It has been reported that four of the five false negatives of the gastric biopsy malpractice claims involve SRCC (25). As for false-positive results, benign signet ring cell changes can mimic SRCC. A collection of "foam cells," which are histiocytes containing phagocytosed mucin or lipid in the lamina propria, is probably the most common histologic entity that resembles SRCC. In our study, we found that one xanthoma (of 10 cases) was misdiagnosed as CA (Supplementary Fig. S3D). Signet ring cell change can also be seen in acute erosive gastropathy in which gastric epithelial cells show signet ring cell changes due to a degenerative process from ischemia (26). As signet ring cell change–containing lesions are rare and sometimes the morphologic features alone are not enough to differentiate between signet ring cell changes and SRCC, further testing with such rare cases is warranted to increase the reliability of our algorithm.

One of the limitations of our algorithm is its inability to detect nonepithelial tumors. Because our model has been trained to recognize only epithelial tumors, it is highly likely that mesenchymal tumors of the stomach, such as GIST, schwannoma, and leiomyoma, would be classified as NFD. As mesenchymal tumors mostly manifest as a submucosal mass beneath the normal gastric mucosa, the biopsy sample tends to contain only a tiny piece of the tumor underneath the muscularis mucosae, requiring careful examination. Indeed, in our prospective study, the algorithm failed to recognize a case of GIST as a tumor and misdiagnosed it as NFD. The majority of mesenchymal tumors require additional IHC testing to make a final diagnosis. Therefore, it may be reasonable to supplement the automated system with a "bypass strategy" so that if a biopsied lesion is described as a submucosal tumor by the endoscopist, the case is categorized separately and sent directly to the pathologist. Interestingly, two neuroendocrine tumors were classified as CA, even though our algorithm was not trained with this type of tumor, which is probably due to its nesting and infiltrative growth patterns that resemble tubular adenocarcinomas.

For countries with a high incidence of gastric cancer where large-scale endoscopic screening is performed on a population basis (such as Japan and Korea), reducing the diagnostic burden on pathologists would be a major benefit anticipated from AI assistance. However, in countries where the incidence of gastric cancer is much lower and diagnostic accuracy is a concern because of a lack of experienced pathologists, our algorithm, which has been trained with a massive number of cases, would also prove useful in detecting gastric cancers. Our algorithm was trained with gastric samples diagnosed by Korean pathologists; thus, there is an important issue to consider when applying the algorithm to specimens from other countries, which involves discrepancies in histologic interpretation of gastric lesions among different countries. For instance, Japan developed diagnostic terminology and criteria for GI epithelial neoplasia different from Western countries, and gastric lesions diagnosed as high-grade dysplasia by most Western pathologists are almost always diagnosed as carcinoma by Japanese pathologists (27). Although the Vienna classification improved diagnostic agreement, in part, by establishing new terminology (17), there still may exist differences in reporting of gastric dysplasia/carcinoma. Korean pathologists have been influenced by both Japanese and Western perspectives, and established diagnostic terminology and guidelines for gastric neoplasia to improve diagnostic consensus (28). Thus, we believe that there would be no significant diagnostic discrepancies when our algorithm is used in other countries; however, it needs verification by pathologists from other countries.

On a final note, in a population with high *Helicobacter pylori* (*H. pylori*) infection rates, such as Korea and Japan, MALT lymphoma has a relatively high prevalence. It was reported that of the 105,194 patients who received screening upper endoscopy from 2003 to 2013 at Seoul National University Hospital (Jongno-gu, Seoul, South Korea), 429 malignancies were detected, of which MALT lymphoma accounted for 12% (51/429 malignancies; ref. 29). These numbers suggest that in countries with high *H. pylori* infection rates, MALT lymphoma should be on the list of gastric tumors for screening. Our test set contained five cases of MALT lymphoma detected and diagnosed by the pathologists. As our algorithm was never trained with cases of MALT lymphoma, four cases (80%) were classified as NFD, and one (20%) was classified as CA due to the destruction of the gastric glands by the lymphoma cells. MALT lymphoma can appear as an overt malignant lesion on endoscopy, but it more often presents as a simple erosion, a thickened gastric fold, a gastritis-like change, or even as normal gastric mucosa (30). For this reason, it is difficult to suspect MALT lymphoma based solely on the endoscopic findings; thus, an "algorithm bypass" strategy that we had suggested for submucosal tumors cannot be applied to MALT lymphomas. High-grade lymphomas, on the other hand, would likely be classified as CA because they are histologically similar to poorly cohesive carcinoma, which at least allows an opportunity for pathologists to review the case.

Published OnlineFirst November 10, 2020; DOI: 10.1158/1078-0432.CCR-20-3159

Therefore, to make our algorithms more complete and practical in the clinical setting as a screening or assistance tool, it is necessary to improve the algorithm to identify MALT lymphoma, in addition to epithelial tumors in gastric biopsies.

In summary, we successfully developed a deep learning algorithm that recognizes and classifies gastric tumors in endoscopic biopsy specimens. In addition to verifying its high accuracy equivalent to experienced human pathologists using a large number of prospectively collected cases, we demonstrated that deep learning algorithm provides substantial time-saving benefits in an assistance mode. Despite several limitations, we believe that our model possesses great potential to serve as a screening or an assistance tool not only in countries with increasing diagnostic workloads for gastric endoscopic specimens, but also in areas where experienced pathologists are not available.

Authors' Disclosures

J. Park reports employment with VUNO Inc. Y.W. Kim reports employment with VUNO Inc. H. Park reports employment with VUNO Inc. B.-h. Kim reports grants from Green Cross Medical Foundation during the conduct of the study. K. Kim reports employment with VUNO Inc. K.-H. Jung reports employment with VUNO Inc and reports being an equity holder of VUNO Inc. No disclosures were reported by the other authors.

Authors' Contributions

J. Park: Conceptualization, formal analysis, writing-review and editing. **B.G. Jang:** Resources, data curation, formal analysis, supervision, writing-review and editing. **Y.W. Kim:** Conceptualization, formal analysis, writing-review and editing. **H. Park:** Project administration. **B.-h. Kim:** Resources, data curation. **M.J. Kim:** Resources, data curation. **H. Ko:** Resources, data curation. **J.M. Gwak:** Resources, data curation, formal analysis. **E.J. Lee:** Project administration. **Y.R. Chung:** Writing-review and editing. **K. Kim:** Formal analysis, writing-review and editing. **J.K. Myung:** Formal analysis. **J.H. Park:** Formal analysis. **D.Y. Choi:** Formal analysis. **C.W. Jung:** Formal analysis. **B.-H. Park:** Formal analysis. **K.-H. Jung:** Conceptualization, formal analysis, supervision, writing-review and editing. **D.-I. Kim:** Conceptualization, resources, data curation, formal analysis, supervision, writing-review and editing.

Acknowledgments

This study was supported by Green Cross Laboratory and VUNO Inc.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received August 12, 2020; revised September 29, 2020; accepted November 4, 2020; published first November 10, 2020.

References

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 2015;136:E359–E86.
2. Sugano K. Screening of gastric cancer in Asia. Best Pract Res Clin Gastroenterol 2015;29:895–905.
3. Miyamoto A, Kuriyama S, Nishino Y, Tsubono Y, Nakaya N, Ohmori K, et al. Lower risk of death from gastric cancer among participants of gastric cancer screening in Japan: a population-based cohort study. Prev Med 2007;44:12–9.
4. Jun JK, Choi KS, Lee H-Y, Suh M, Park B, Song SH, et al. Effectiveness of the Korean National Cancer Screening Program in reducing gastric cancer mortality. Gastroenterol 2017;152:1319–28.
5. Suh M, Choi KS, Park B, Lee YY, Jun JK, Lee D-H, et al. Trends in cancer screening rates among Korean men and women: results of the Korean National Cancer Screening Survey, 2004–2013. Cancer Res Treat 2016;48:1–10.
6. Biscotti CV, Dawson AE, Dziura B, Galup L, Darragh T, Rahemtulla A, et al. Assisted primary screening using the automated ThinPrep imaging system. Am J Clin Pathol 2005;123:281–7.
7. Bejnordi BE, Veta M, Van Diest PJ, Van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017;318:2199–210.
8. Komura D, Ishikawa S. Machine learning methods for histopathological image analysis. Comput Struct Biotechnol J 2018;16:34–42.
9. Han Z, Wei B, Zheng Y, Yin Y, Li K, Li S. Breast cancer multi-classification from histopathological images with structured deep learning model. Sci Rep 2017;7:4172.
10. Oikawa K, Saito A, Kiyuna T, Graf HP, Cosatto E, Kuroda M. Pathological diagnosis of gastric cancers with a novel computerized analysis system. J Pathol Inform 2017;8:5.
11. Yoshida H, Shimazu T, Kiyuna T, Marugame A, Yamashita Y, Cosatto E, et al. Automated histological classification of whole-slide images of gastric biopsy specimens. Gastric cancer 2018;21:249–57.
12. Sharma H, Zerbe N, Klempert I, Hellwich O, Hufnagl P. Deep convolutional neural networks for automatic classification of gastric carcinoma using whole slide images in digital histopathology. Compu Med Imaging Graph 2017;61:2–13.
13. Iizuka O, Kanavati F, Kato K, Rambeau M, Arihiro K, Tsuneki M. Deep learning models for histopathological classification of gastric and colonic epithelial tumours. Sci Rep 2020;10:1504.
14. Wang S, Zhu Y, Yu L, Chen H, Lin H, Wan X, et al. RMDL: recalibrated multi-instance deep learning for whole slide gastric image classification. Med Image Anal 2019;58:101549.
15. Song Z, Zou S, Zhou W, Huang Y, Shao L, Yuan J, et al. Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning. Nat Commun 2020;11:4294.
16. Li Y, Li X, Xie X, Shen L. Deep learning based gastric cancer identification. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI, April 4–7 2018).
17. Schlemper R, Riddell R, Kato Y, Borchard F, Cooper H, Dawsey S, et al. The Vienna classification of gastrointestinal epithelial neoplasia. Gut 2000;47:251–5.
18. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115–8.
19. Litjens G, Sánchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep 2016;6:26286.
20. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nat med 2018;24:1559–67.
21. Steiner DF, MacDonald R, Liu Y, Truszkowski P, Hipp JD, Gammage C, et al. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. Am J Surg Pathol 2018;42:1636–46.
22. Levine AB, Schlosser C, Grewal J, Coope R, Jones SJ, Yip S. Rise of the machines: advances in deep learning for cancer diagnosis. Arch Pathol Lab Med 2006;130:617–9.
23. Tatsuta M, Iishi H, Okuda S, Oshima A, Taniguchi H. Prospective evaluation of diagnostic accuracy of gastrofiberscopic biopsy in diagnosis of gastric cancer. Cancer 1989;63:1415–20.
24. Kloeckner J, Sansonowicz TK, Rodrigues ÁL, Nunes TWN. Multi-categorical classification using deep learning applied to the diagnosis of gastric cancer. J Bras Pathol Med Lab 2020;56:e1522020.
25. Troxel DB. Medicolegal aspects of error in pathology. Arch Pathol Lab Med 2006;130:617–9.
26. Dimet S, Lazure T, Bedossa P. Signet-ring cell change in acute erosive gastropathy. Am J Surg Pathol 2004;28:1111–2.
27. Schlemper RJ, Kato Y, Stolte M. Diagnostic criteria for gastrointestinal carcinomas in Japan and Western countries: proposal for a new classification system of gastrointestinal epithelial neoplasia. J Gastroenterol Hepatol 2000;15:G49–G57.
28. Kim JM, Cho M-Y, Sohn JH, Kang DY, Park CK, Kim WH, et al. Diagnosis of gastric epithelial neoplasia: dilemma for Korean pathologists. World J Gastroenterol 2011;17:2602.
29. Yang HJ, Lee C, Lim SH, Choi JM, Yang JI, Chung SJ, et al. Clinical characteristics of primary gastric lymphoma detected during screening for gastric cancer in Korea. J Gastroenterol Hepatol 2016;31:1572–83.
30. Zullo A, Hassan C, Ridola L, Repici A, Manta R, Andriani A. Gastric MALT lymphoma: old and new insights. Ann Gastroenterol 2014;27:27–33.

AACR American Association for Cancer Research

Clinical Cancer Research

A Prospective Validation and Observer Performance Study of a Deep Learning Algorithm for Pathologic Diagnosis of Gastric Tumors in Endoscopic Biopsies

Jeonghyuk Park, Bo Gun Jang, Yeong Won Kim, et al.

Clin Cancer Res 2021;27:719-728. Published OnlineFirst November 10, 2020.