ORIGINAL ARTICLE

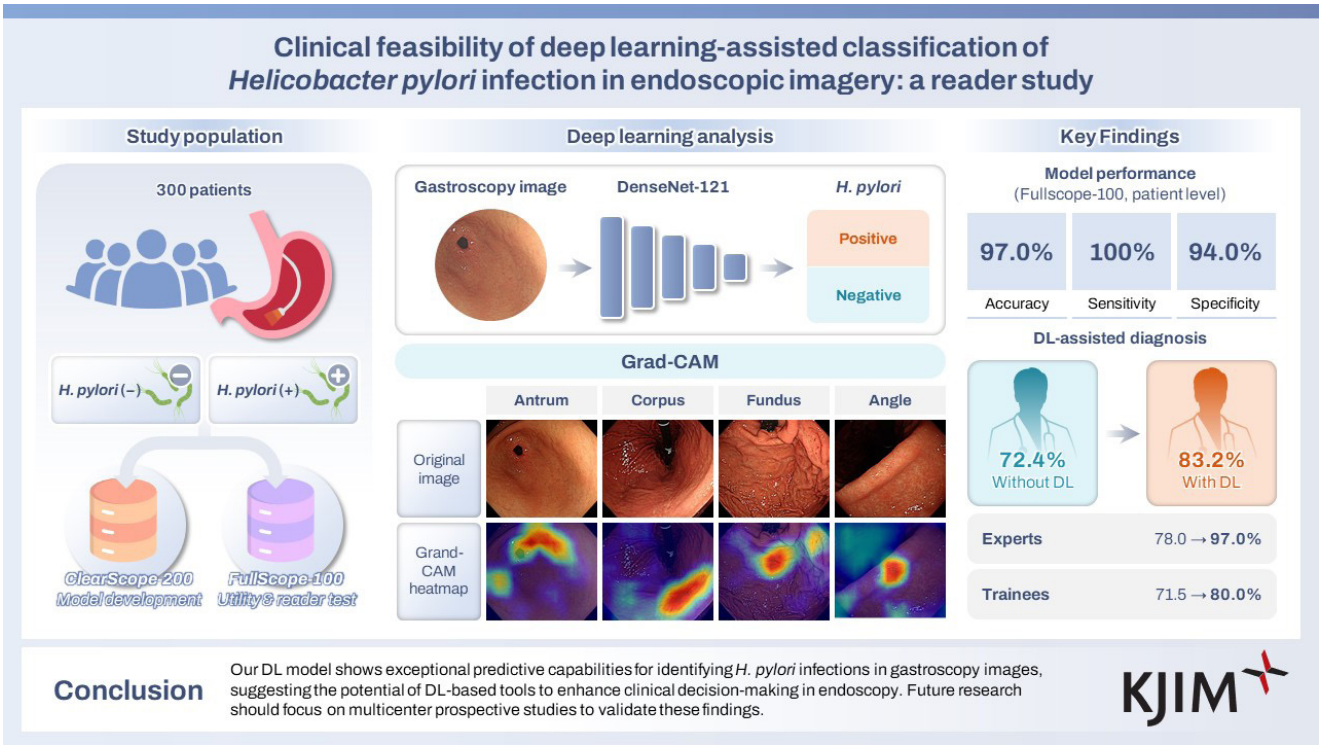
Korean J Intern Med 2026;41:862-872
https://doi.org/10.3904/kjim.2024.245

Clinical feasibility of deep learning-assisted classification of *Helicobacter pylori* infection in endoscopic imagery: a reader study

Jun-young Seo[1,2,*], Jiseon Kang[3,*], Do Hoon Kim[1], and Namkug Kim[4]

[1]Department of Gastroenterology, University of Ulsan College of Medicine, Asan Medical Center, Seoul; [2]Digestive Disease Center, CHA Bundang Medical Center, CHA University School of Medicine, Seongnam; [3]Department of Medicine, University of Ulsan College of Medicine, Asan Medical Center, Seoul; [4]Department of Convergence Medicine, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Korea

*These authors contributed equally to this manuscript.



Background/Aims: Detecting *Helicobacter pylori* infection through endoscopic imaging is a preferred method to address the limitations of invasive diagnostics. We have developed a deep learning (DL) model to classify these images based on their *H. pylori* infection status and to evaluate its potential as a diagnostic aid.

Methods: We retrospectively enrolled *H. pylori*-positive patients (by rapid urease test and serum IgG) at Asan Medical Center between January 2015 and December 2020, establishing development and test datasets to evaluate model utility. We developed a DL model and assessed its performance at both the image and patient levels using metrics such as accuracy, sensitivity, specificity, and F1-score. Fourteen readers of varying experience levels interpreted the endoscopic images in two

pISSN 1226-3303
eISSN 2005-6648
http://www.kjim.org

sessions, before and after the integration of the DL model results. The diagnostic accuracy of the readers was analyzed using the McNemar test and generalized estimating equations.

Results: In the utility test set, the image-level accuracy, sensitivity, specificity, and F1-score of our DL model were 84.2%, 74.4%, 93.9%, and 82.5%, respectively. At the patient level, the corresponding values were 97.0%, 100%, 94.0%, and 97.1%. Endoscopists utilizing the DL model demonstrated significantly improved accuracy in classifying *H. pylori* infections, with classification rates of 83.2% versus 72.4% ($p < 0.001$).

Conclusions: Our DL model shows exceptional predictive capabilities for identifying *H. pylori* infections in gastroscopy images, suggesting the potential of DL-based tools to enhance clinical decision-making in endoscopy. Future research should focus on multicenter prospective studies to validate these findings.

Keywords: Artificial intelligence; Deep learning; Endoscope; *Helicobacter pylori*

INTRODUCTION

Helicobacter pylori, first identified in the 1980s, now exhibits a global prevalence exceeding 50% [1,2]. Given its association with several gastric diseases, including atrophic gastritis, intestinal metaplasia, peptic ulcer, and gastric cancer, accurate diagnosis and eradication of *H. pylori* infection are critical [3].

Various diagnostic approaches are available for detecting *H. pylori*. These methods are categorized into invasive and non-invasive tests. Non-invasive tests predominantly include the urea breath test and serological testing. The urea breath test, with sensitivity and specificity around 95%, may yield false negatives in patients using proton pump inhibitors (PPIs) or those experiencing gastric bleeding, potentially compromising test accuracy [4,5]. Serological tests, while useful, struggle to differentiate between active and previous infections, as they can remain positive for years following successful treatment [6,7]. In cases where endoscopy is planned, several invasive tests such as biopsy, rapid urease test (RUT), *H. pylori* polymerase chain reaction (PCR), and culture are available. However, the uneven distribution of *H. pylori* across the gastric mucosa can lead to false negatives if the bacteria are absent in the tested area [8,9]. Additionally, these invasive tests may heighten the risk of bleeding, particularly in patients on antithrombotic therapy or those with conditions like liver cirrhosis that predispose to bleeding [10]. Nevertheless, diagnosing *H. pylori* based solely on endoscopic findings could potentially reduce healthcare costs by eliminating unnecessary tests and mitigating the risks associated with invasive diagnostic methods.

For decades, artificial intelligence (AI), especially when employing deep learning (DL) technologies based on convolutional neural networks (CNNs), has shown significant promise in medical diagnosis and disease detection. These DL methods are particularly effective in analyzing medical data and identifying patterns indicative of bacterial presence. Meta-analyses have demonstrated that CNN-based computer-aided diagnosis systems for *H. pylori* infection achieve high diagnostic accuracy, confirming the reliability of AI algorithms in endoscopic diagnosis [11,12]. Recent research on *H. pylori* infection has primarily focused on developing accurate and efficient models, culminating in the creation of methods that enhance performance through the integration of advanced DL classification models [13]. Moreover, a study introduced a modified DL model incorporating cutting-edge CNN modules to detect atrophic gastritis caused by *H. pylori* infection [14]. However, these studies often involved limited model search and validation, and did not explore the practical benefits for endoscopists in diagnosing *H. pylori*.

In this study, our objectives were to develop a DL model that accurately categorizes endoscopic images based on *H. pylori* infection status and to assess its utility for endoscopists. We specifically evaluated the accuracy of gastroscopic diagnosis of *H. pylori* infection by comparing the performance of endoscopists assisted by our DL model with that of unassisted endoscopists, thereby exploring the potential clinical applications of our computer-aided diagnostic system.

METHODS

The institutional review board of Asan Medical Center ap-

proved this study protocol (IRB No. 2020-1574). The requirement for informed consent was waived by the board.

Data set and ground truth

We conducted a retrospective collection of data from patients who underwent screening endoscopy at the Health Screening and Promotion Center of Asan Medical Center between January 2015 and December 2020. All patients underwent the RUT and a serum *H. pylori*-specific IgG antibody test. The RUT, provided by Chongkeundang Pharma in Seoul, Korea, involved assessing two gastric specimens from the antrum and body. Each specimen was placed into separate RUT kits, labeled A and B. A negative RUT outcome was indicated by no color change within 24 hours. We identified individuals as *H. pylori*-positive if their RUT result was positive and their IgG serology levels were 8 or above. Conversely, patients were classified as non-infected if their RUT result was negative and their IgG serology level was 0. We excluded all patients with intermediate IgG values. The endoscopic images of selected patients were retrieved from the Picture Archiving and Communication System (PACS). These images had been acquired following a standardized imaging protocol, which ensured capture from the same anatomical locations (antrum, angle, body, and fundus of the stomach). Through this protocol, an average of 13 images (ranging from 8 to 18) were collected per patient using a GIF-290HQ gastroscope (Olympus, Tokyo, Japan). They were used to document the presence or absence of *H. pylori* infection based on the white light imaging findings.

Data partitions

Our study included a total of 300 patients, comprised of 150 *H. pylori*-negative and 150 *H. pylori*-positive individuals, contributing 1,561 *H. pylori*-negative images and 1,178 *H. pylori*-positive images. We organized the collected data into two temporally distinct datasets based on individual patients: ClearScope-200 and FullScope-100. ClearScope-200 included 200 patients—100 *H. pylori*-negative and 100 *H. pylori*-positive. We excluded images that hindered diagnosis to facilitate the development of a DL model, subsequently dividing this dataset into training, validation, and test sets. Conversely, FullScope-100 consisted of 100 patients, with equal numbers of *H. pylori*-negative and -positive cases, encompassing all collected images. This unfiltered dataset provides a realistic view of clinical data, enabling us to evaluate the model's practical performance. We employed the FullScope-100 dataset for a reader study to examine the effects of DL support on physician decision-making. The distribution of images across the data splits is detailed in Supplementary Table 1.

DL model for endoscopic image classification

The endoscopic images sourced from the PACS were labeled with text, which could significantly hinder model training. To address this, preprocessing was conducted to isolate the actual endoscopic image area of 466 × 410 pixels (see Supplementary Fig. 1). To prevent overfitting due to limited data, various data augmentation techniques were implemented, including random flipping, rotation, color jittering, sharpening, Gaussian blurring, and elastic transformation. Moreover, to counter the skewed distribution of *H. pylori*-negative and *H. pylori*-positive images, a weight-based random sampling method was employed during dataset compilation. This method assigns higher sampling probabilities to the underrepresented class and lower to the overrepresented, thus balancing the dataset.

For the classification of endoscopic images as *H. pylori*-positive or not, the DenseNet-121 [15] CNN was used, renowned for its effectiveness in image classification tasks. This network comprises 121 layers and features a unique architecture in which each layer connects directly to every other layer in a feed-forward manner. This design allows the network to be both deeper and more efficient, without an increase in computational complexity, and facilitates the processing of input images at their original size. The final fully connected (FC) layer of the network was modified to produce a binary classification output. All layers, except the final FC layer, were initialized with pretrained weights from the ImageNet [16] dataset, and the model was trained on 32-sized image batches over 150 epochs. The learning rate was adjusted dynamically: starting at 0.01 for the initial 50 epochs, decreasing to 0.001 for the subsequent 50 epochs, and further reducing to 0.0001 for the final 50 epochs. Cross-entropy loss was employed as the loss function, and Stochastic Gradient Descent with a momentum of 0.9 was used as the optimizer. The output logits from the model were processed using the sigmoid function to determine the probability of an image being *H. pylori* negative or positive. Concurrently, a Gradient-Weighted Class Activation Map (Grad-CAM) [17] was generated to visually highlight the regions deemed significant by the model in making predictions, based on the gradient information from the last

https://doi.org/10.3904/kjim.2024.245

convolutional layer. The algorithm was implemented using Python (version 3.8.15) and the PyTorch framework (version 1.13.1). All experiments were conducted on a workstation equipped with dual Intel Xeon CPU E5-2630 v4, each with 10 cores and a clock speed of 2.2 GHz, 128GB of DDR RAM, and an NVIDIA TITAN RTX GPU with 24GB of GDDR6 memory.

Reader accuracy evaluation

The observer panel comprised 14 physicians, including two experienced endoscopists and 12 gastrointestinal fellows, each with varying levels of gastroscopy interpretation experience. The two endoscopists had over three years of clinical experience in identifying *H. pylori*, while the fellows each had less than one year of experience. A reader test was conducted using the FullScope-100 dataset in two phases: DL-unassisted and DL-assisted interpretations (Supplementary Fig. 2). This test, a paired design with sequential sessions, was chosen for its suitability for evaluating our DL model for medical diagnosis [18]. In the initial session, readers interpreted endoscopic images without DL assistance. Four weeks later, in the subsequent session, the same readers re-evaluated the images with DL assistance, which included Grad-CAM visualizations for each image and patient-level model predictions. All readers were blinded to both the diagnosis and clinical background information.

Statistical analysis

The performance of our DL model was evaluated using metrics such as accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1-score across two test datasets. The overall classification efficacy was further determined by plotting receiver operating characteristic (ROC) curves and calculating the area under the curve (AUC) score. Image-level results were aggregated per patient, with a patient classified as *H. pylori*-positive if more than 25% of their images tested positive; specifically, this was operationally defined as a confirmed infection in at least 2 to 4 of the 8–18 standard anatomical images captured per patient. This threshold was set acknowledging the diffuse nature of *H. pylori* colonization in the stomach [19].

The impact of DL assistance on endoscopists' diagnostic capabilities was assessed using the McNemar test, analyzing changes in diagnostic accuracy before and after the introduction of DL assistance. To account for the clustered Multi-Reader Multi-Case design, data were stratified by the experience level of the readers, and statistical differences were evaluated using a logistic regression model with generalized estimating equations (GEE) across different reader groups and the entire dataset. Analyses were performed using statistical software (R version 4.3.1, R Project for Statistical Computing), with p values less than 0.05 deemed statistically significant.

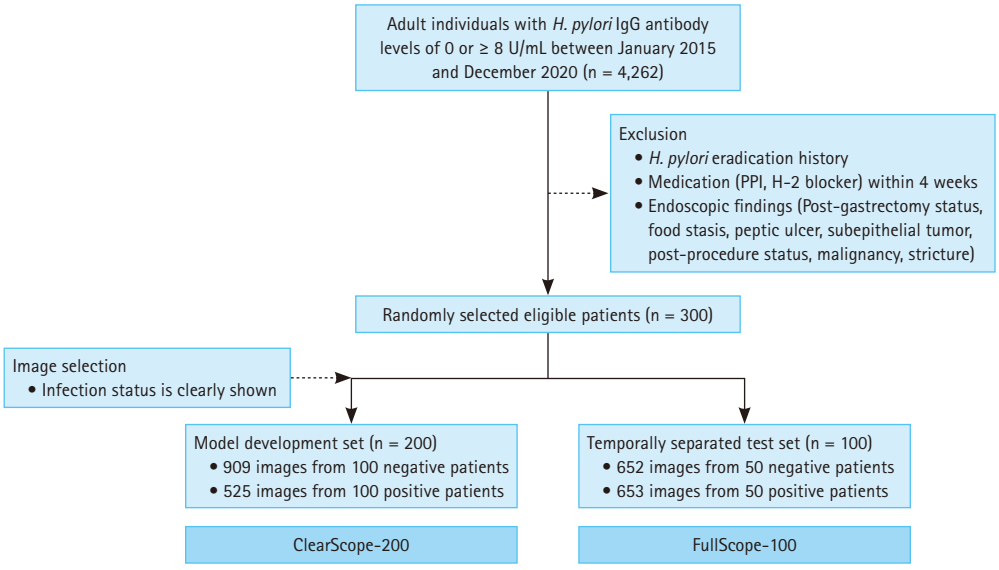


Figure 1. Flowchart of the study. *H. pylori*, *Helicobacter pylori*; PPI, proton pump inhibitor.

Table 1. Performance of our deep learning-based *Helicobacter pylori* classification model

Matric	ClearScope-200 test set (20 patients; 131 images)		FullScope-100 (100 patients; 1,305 images)	
	Image-Level	Patient-Level	Image-Level	Patient-Level
Accuracy	89.5 (84.0, 93.9)	94.3 (80.0, 100.0)	84.2 (82.3, 86.0)	97.0 (94.0, 100.0)
Sensitivity	85.5 (73.1, 94.3)	100.0 (100.0, 100.0)	74.4 (71.1, 77.6)	100.0 (100.0, 100.0)
Specificity	91.7 (85.5, 96.6)	88.7 (66.4, 100.0)	93.9 (91.9, 95.5)	94.0 (87.5, 100.0)
PPV	85.2 (75.0, 94.5)	89.5 (64.7, 100.0)	92.5 (90.2, 94.5)	94.4 (88.5, 100.0)
NPV	91.8 (85.7, 96.9)	100.0 (100.0, 100.0)	78.6 (76.1, 81.7)	100.0 (100.0, 100.0)
F1-score	85.2 (75.9, 92.0)	94.1 (78.6, 100.0)	82.5 (80.2, 84.6)	97.1 (93.9, 100.0)

The table displays results on two test sets: the test split of the ClearScope-200 dataset (20 patients; 131 images) and the FullScope-100 test set (100 patients; 1,305 images). The "Image-Level" columns display results for each image, while the "Patient-Level" columns aggregate these results for each patient, with a patient being classified as *H. pylori*-positive if more than 25% of images are classified as positive. Data are presented in percentages with 95% confidence intervals.
PPV, positive predictive value; NPV, negative predictive value; F1-score, the harmonic mean of the precision and recall.

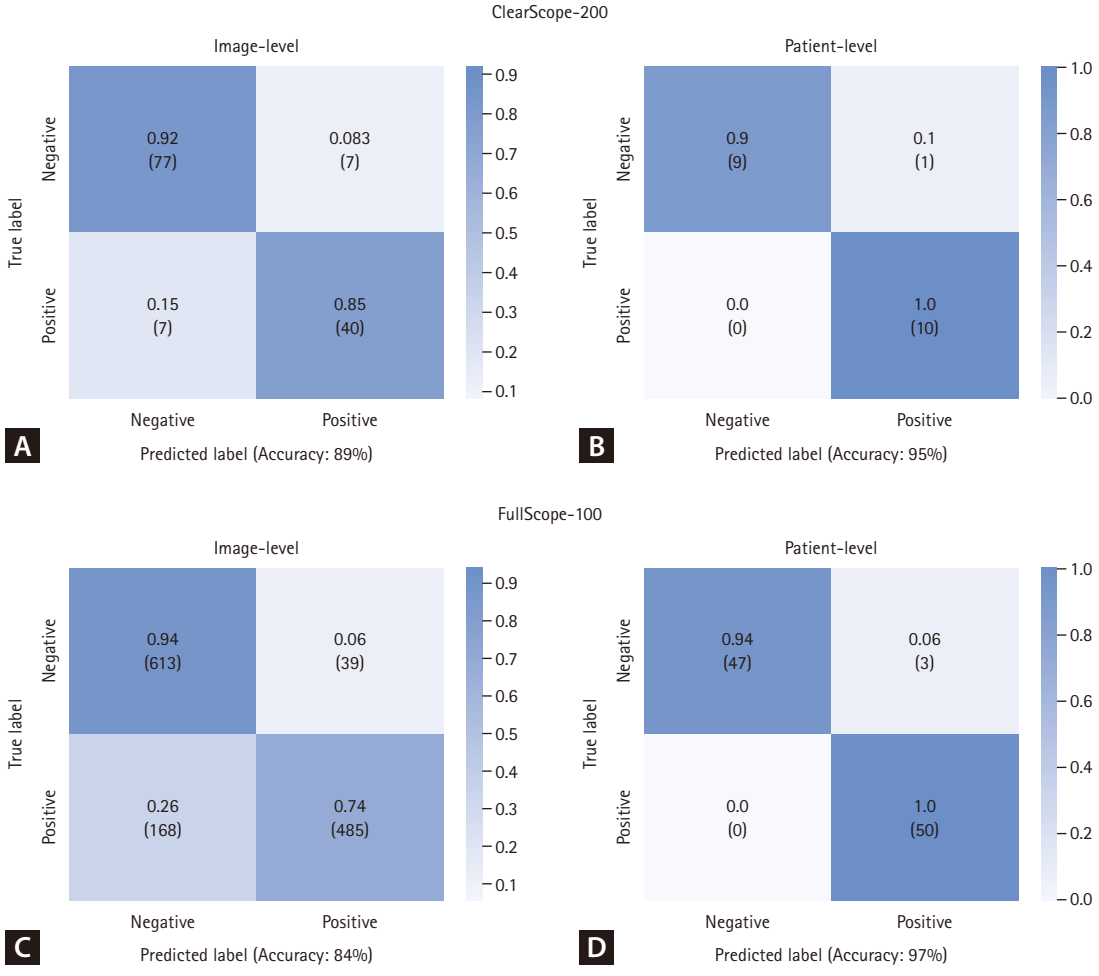


Figure 2. Confusion matrices of our deep learning model on the test sets of ClearScope-200 and FullScope-100. The first row presents results from the ClearScope-200 test set at both the image level (A) and the patient level (B). The second row shows results from the FullScope-100 test set at the image level (C) and the patient level (D). The matrices are color-coded and values are based on normalized data. Numbers in parentheses indicate the count of images and patients at the image and patient levels, respectively.

https://doi.org/10.3904/kjim.2024.245

RESULTS

Patient characteristics

Among a cohort of 4,262 patients (2,507 *H. pylori*-negative and 1,755 *H. pylori*-positive), we excluded those with endoscopic findings of post-gastrectomy stomach, food stasis, peptic ulcer, subepithelial tumor, malignancy, post-procedure status, or stricture. Additionally, patients with a history of *H. pylori* eradication treatment and those who had taken PPIs or H2 blockers within 4 weeks prior to endoscopy were also excluded following a review of medical records. After these exclusions, a total of 300 patients (150 *H. pylori*-negative, 150 *H. pylori*-positive) were randomly selected and divided into two datasets for model development and independent testing (Fig. 1). The model development set, ClearScope-200, included 1,434 endoscopic images from 200 patients (mean age, 57 years ± 10.4 [standard deviation]; 103 males), and the temporally separate test set, FullScope-100, comprised 1,305 images from 100 patients (mean age, 53.5 years ± 9.6; 57 males) (Supplementary Table 2).

Model performance

The performance of our DL model was assessed using two test datasets: the ClearScope-200 and the FullScope-100 test sets. The ClearScope-200 test set underwent the same image curation process as the training set, whereas the FullScope-100 test set included a broader array of images. We computed the mean performance metrics and their corresponding 95% confidence intervals (CI) using the bootstrap method (Table 1). For the ClearScope-200 test set, the model demonstrated an image-level accuracy of 89.5%, sensitivity of 85.5%, specificity of 91.7%, PPV of 85.2%, NPV of 91.8%, and F1-score of 85.2. At the patient-level, these metrics were 94.3% for accuracy, 100.0% for sensitivity, 88.7% for specificity, 89.5% for PPV, 100.0% for NPV, and 94.1% for F1-score, respectively. For the FullScope-100 test set, the image-level metrics were 84.2% for accuracy, 74.4% for sensitivity, 93.9% for specificity, 92.5% for PPV, and 82.5% for NPV and F1-score. At the patient-level, the metrics improved to 97.0% for accuracy, 100.0% for sensitivity, 94.0% for specificity, 94.4% for PPV, 100.0% for NPV, and 97.1% for F1-score. Overall, performance metrics improved when image-level results were aggregated to patient units, as evidenced by the reduction in false negative instances displayed in the confusion matrices in Figure 2.

Our DL model achieved 100% patient-level sensitivity across both test cohorts. It is important to address that this perfect sensitivity is not an indication of model overfitting, but rather reflects the robustness of the patient-level aggregation method. First, the model demonstrated imperfect image-level sensitivity (74.4–85.5%), confirming it was not

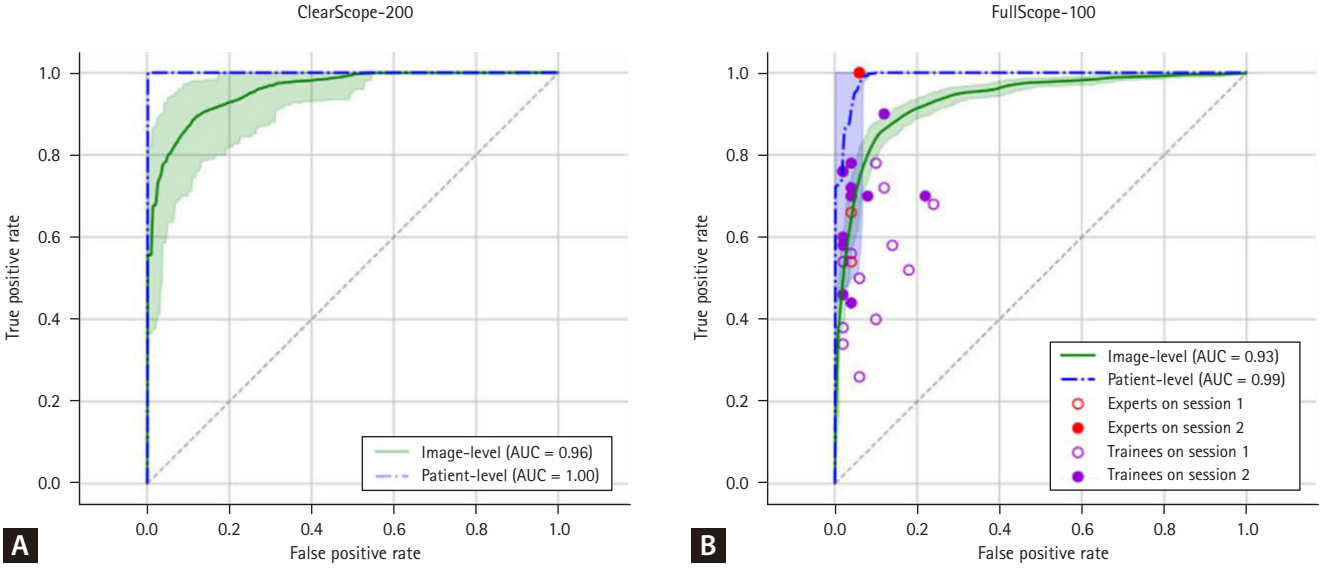


Figure 3. Receiver operating characteristic (ROC) curves of the deep learning model with 95% confidence intervals. This figure displays ROC curves at the image level (green) and the patient level (blue). Panel A illustrates performance on the ClearScope-200 test set, while Panel B shows performance on the FullScope-100 test set, including physician endoscopist results at the patient level. AUC, area under the curve.

overfit to individual images. Second, this 100% sensitivity was replicated on the FullScope-100 set, a temporally separated, independent cohort, which indicates strong generalization. This perfect sensitivity is achieved because the aggregation rule (patient-positive if > 25% of images are positive) is highly resilient to image-level errors. For a patient-level 'False Negative' to occur, the model would need to fail on more than 75% of the images from a positive patient. Given the model's high image-level accuracy, the probability of such a compounded failure is minimal, resulting in the observed 100% sensitivity.

Figure 3 illustrates the ROC curves for both test sets, providing insight into the model's ability to distinguish between negative and positive classes. Higher AUC scores indicate superior discrimination, with scores of 0.96 and 1.00 for the ClearScope-200 test set at the image-level and patient-lev-

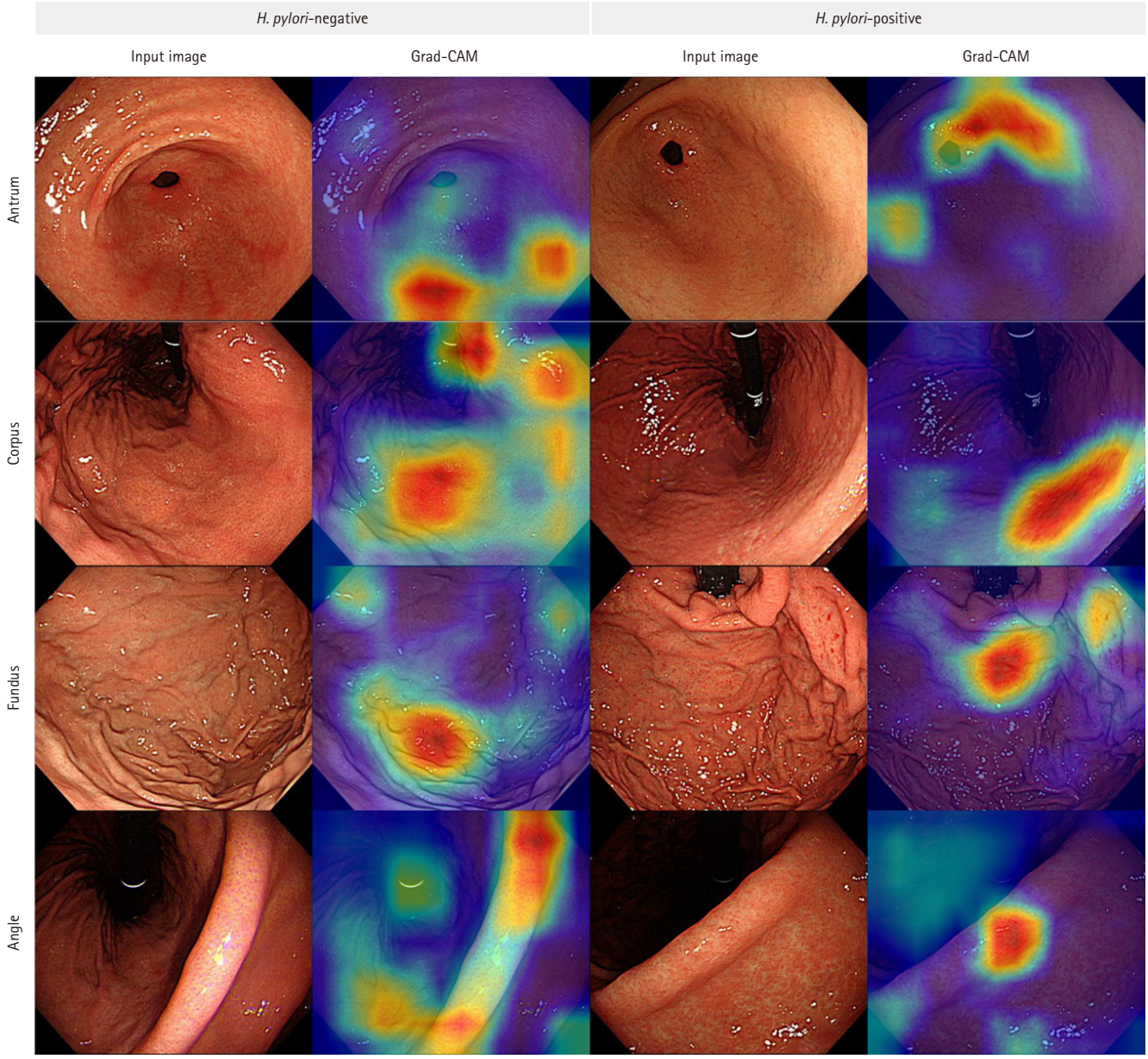


Figure 4. Gradient-weighted Class Activation Mapping (Grad-CAM) visualizations. This figure uses heatmap visualization to highlight regions deemed important by the deep learning model. The leftmost pair of columns presents endoscopic images alongside their corresponding Grad-CAM visualizations for control patients. The rightmost pair of columns displays endoscopic images and their associated Grad-CAM visualizations for patients diagnosed with *Helicobacter pylori* infection.

https://doi.org/10.3904/kjim.2024.245

el, respectively, and 0.93 and 0.99 for the FullScope-100 test set. These findings demonstrate that the discriminative ability of our DL model is consistently robust across both datasets, regardless of the level of analysis. Notably, the ClearScope-200 test set showed a wider CI range for the ROC curve compared to the FullScope-100 test set, likely reflecting the smaller sample size and fewer images per patient in the ClearScope-200 dataset, as discussed in Table 1.

The Grad-CAM technique highlights the specific areas of an image that the model prioritizes when predicting a given class. Values typically range from 0 to 1, with higher values indicating critical regions for class prediction. These are visualized in a heatmap, where areas of high importance are colored red and areas of low importance are colored blue. Figure 4 presents representative Grad-CAM images for our model. The rows in each figure represent the antrum, corpus, fundus, and angle of the stomach, respectively. The second column's Grad-CAM heatmap corresponds to the *H. pylori*-negative image in the first column, indicating the absence of *H. pylori* infection. Conversely, the heatmap in the fourth column aligns with the *H. pylori*-positive image in the third column, marking the presence of *H. pylori* infection through visual analysis.

Diagnostic accuracy of readers

The results from the patient-level reader test, which assessed the diagnostic impact of DL support across 14 participants, are detailed in Table 2. In Session 1, without DL assistance, the average diagnostic accuracy for all participants was 72.4% (95% CI: 68.8–76.0%). With the integration of our DL model in Session 2, providing patient-level predictions and corresponding Grad-CAM visualizations for each image, accuracy improved to 83.2% (95% CI: 78.9–87.5%). This improvement was found to be highly statistically significant. The GEE analysis yielded a Z-value of 6.984 (p = 2.87e-12). Furthermore, a post-hoc power analysis based on this observed result confirmed that the statistical power (1-β) of our study was > 99.9%, demonstrating that the study was robustly powered to detect this effect despite the N = 100 patient cohort. Individual reader results are depicted in the ROC curves of our DL model (Fig. 3B), with outlined dots for Session 1 and filled dots for Session 2 results.

Accuracy improved across all reader groups, although some variations in performance characteristics were noted. Experts with over three years of endoscopy experience demonstrated a significant increase in the accuracy of diagnosing *H. pylori* after receiving DL support, with diagnostic accuracy improving from 78.0% before DL assistance to 97.0% afterward (p < 0.001). The accuracy of our DL model reached 97.0% on the FullScope-100 dataset, matching the performance of these experts in Session 2 with DL assistance. Their decision-making, influenced by Grad-CAM outputs and the displayed *H. pylori* infection status, suggests that Grad-CAM effectively guided the experts in their diagnoses. Conversely, while trainees also showed significant improvements in classifying *H. pylori* gastritis with DL support—increasing from 71.5% to 80.0% accuracy (p < 0.001)—the

Table 2. Patient-level diagnostic accuracy of physician endoscopists (2 experts, 12 trainees) on the FullScope-100 test set (100 patients; 1,305 images), with and without deep learning (DL) assistance

Parameter	Session 1 (without DL assistance)	Session 2 (with DL assistance)	p value
Experts			
Reader 1	81.0	97.0	< 0.001[b]
Reader 2	75.0	97.0	< 0.001[b]
Group mean (Readers 1-2)	78.0 (72.1, 83.9)[a]	97.0 (97.0, 97.0)[a]	< 0.001[c]
Trainees			
Reader 3	67.0	87.0	< 0.001[b]
Reader 4	72.0	74.0	0.860[b]
Reader 5	66.0	79.0	0.007[b]
Reader 6	76.0	83.0	0.007[b]
Reader 7	80.0	89.0	0.035[b]
Reader 8	72.0	84.0	0.023[b]
Reader 9	60.0	70.0	0.013[b]
Reader 10	68.0	72.0	0.344[b]
Reader 11	84.0	87.0	0.607[b]
Reader 12	65.0	81.0	0.001[b]
Reader 13	72.0	87.0	< 0.001[b]
Reader 14	76.0	78.0	0.804[b]
Group mean (Readers 3-14)	71.5 (67.7, 75.3)[a]	80.0 (77.3, 84.5)[a]	< 0.001[c]
Overall mean (Readers 1-14)	72.4 (68.8, 76.0)[a]	83.2 (78.9, 87.5)[a]	< 0.001[c]

[a]Data in parentheses are 95% confidence intervals for the group mean or overall mean.
[b]p value was calculated using the McNemar test.
[c]p value was calculated using a logistic regression model with generalized estimating equations.

extent of their improvement was less pronounced than that observed among the experts. Moreover, the increase in diagnostic accuracy among certain trainees (Reader 4, 10, 11, and 14) was not statistically significant, possibly indicating that these individuals might have correctly identified challenging samples by chance. This observation suggests that the trainees' limited ability to interpret Grad-CAM visualizations may stem from inadequate training in recognizing the varied endoscopic features of *H. pylori* gastritis.

DISCUSSION

To accurately diagnose active *H. pylori* infections, clinicians often utilize invasive methods during gastroscopy, which are associated with significant costs, time demands, and the potential for complications. In response to these challenges, there has been increasing research interest in the prediagnosis of *H. pylori* infections using endoscopic imagery [20]. While various diagnostic tools exist to detect *H. pylori* infections [21], it is not clinically feasible to employ multiple tests simultaneously due to their individual limitations in accuracy. In this context, employing a DL model such as ours is expected to enhance the accuracy of *H. pylori* diagnosis and offer cost-effective benefits to patients.

Recent meta-analyses have shown that AI diagnosis of *H. pylori* gastritis yields a ROC of 0.92 (95% CI 0.90–0.94). However, these studies were conducted retrospectively and did not provide evidence of their utility in clinical practice [12]. In contrast, our study demonstrates that a DL-based image classification model can accurately classify *H. pylori* infections and assist endoscopists in diagnosis. We trained our DL model on endoscopic images with confirmed *H. pylori* infections and evaluated it on a dataset collected according to standard endoscopic protocols. Our model achieved an accuracy of 84.2% in classifying images from this unrefined dataset. Additionally, when analyzing results on a per-patient basis, our model identified the presence of *H. pylori* infection with 97% accuracy. These findings are comparable to a recent study using ShuffleNet, a well-established CNN, which achieved an atrophic sensitivity of 98.5% and a normal sensitivity of 97.9% in detecting *H. pylori* gastritis in the antrum region of the stomach [14]. Given that our model was trained and evaluated across all gastric regions, it can be considered more robust in various gastric environments.

Using DL techniques to detect *H. pylori* in endoscopic images is not a novel concept. Prior applications of DL in endoscopy have shown considerable promise in identifying *H. pylori* infection status [13]. However, the majority of studies have focused primarily on enhancing the accuracy of endoscopic image classifiers and have not thoroughly assessed their practical utility for endoscopists. In this study, we developed a DL system that aids the diagnostic process by providing physicians with patient-level diagnoses and image-level saliency maps (Grad-CAM heatmaps), thereby demonstrating the model's potential for clinical application in the diagnosis of *H. pylori* infection.

In a panel consisting of both inexperienced trainees and expert endoscopists, the implementation of DL assistance significantly improved diagnostic accuracy (from 72.4% to 83.2% with DL assistance; $p < 0.001$). Our findings indicate that DL technology can substantially enhance the accuracy of *H. pylori* infection diagnosis, which in turn could greatly benefit endoscopists in classifying the infection, regardless of their experience level. Moreover, experts with an in-depth understanding of both the positive and negative signs of *H. pylori* found the inclusion of Grad-CAM images as saliency maps particularly helpful (as shown in Table 2). In contrast, although diagnostic accuracy improved for all trainees in the second session, for four trainees (Readers 4, 10, 11, and 14), the improvement was not statistically significant. This may have been due to the potential confusion caused by the Grad-CAM images, given their limited experience with endoscopic signs of *H. pylori*. Nevertheless, across all participants, the average diagnostic accuracy in the second session showed a significant improvement over the first session, suggesting that assistance from the DL model was likely beneficial to endoscopists overall.

Our study has several limitations. First, distinguishing true non-infection from past (eradicated) infection was challenging. Our 'non-infected' group, defined by chart review, may still include individuals with unintended eradication from other antibiotics [22]. Future studies should incorporate serological assessments, such as the pepsinogen test, to definitively identify past infections. Second, selecting patients from both the positive and negative cohorts could introduce selection bias. This problem often arises during the selection of suitable patient cohorts for model development and validation at a single center, particularly with a small sample size. Third, the collection of endoscopic images from various gastric regions, rather than a specific area, complicates the comparison of performance across these regions. According

https://doi.org/10.3904/kjim.2024.245

to the recent Kyoto classification of gastritis, different endoscopic manifestations of *H. pylori* gastritis are defined based on the regions of the stomach [23]. Analyzing diagnostic accuracy by specific gastric regions could provide clinicians with more precise diagnostic information and aid in the advancement of more effective DL models. Fourth, our reader study used a sequential paired design. While our 4-week washout period and case randomization aimed to mitigate recall bias, a different pitfall remains [18]. Although case order was randomized and clinical information was blinded to minimize recall bias, the possibility of residual recall or learning effects cannot be completely excluded. In particular, readers may have become more familiar with the interpretation framework during the second session, which could have contributed to improved performance independent of DL assistance. Finally, the reader panel was imbalanced, with a small number of experts compared to many trainees. While this composition aimed to capture the wide variability in trainee diagnostic skills against a stable expert baseline, it could potentially skew the overall mean values toward the more numerous trainee group. We performed a subgroup analysis stratified by experience level (Table 2) to address this potential bias.

In conclusion, our study underscores the promise of DL-assisted diagnosis for *H. pylori* infection. The proposed DL-based diagnostic system shows potential for direct clinical integration, enabling early detection of *H. pylori* during gastroscopy, independent of the endoscopist's expertise. By providing real-time feedback on the likelihood of *H. pylori* infection and visual heatmap guidance, this system could help endoscopists identify suspicious mucosal areas and prioritize biopsy sites more efficiently. Furthermore, such DL assistance may reduce unnecessary testing and improve diagnostic confidence, particularly among less-experienced endoscopists. Beyond diagnostic support, this model could also be applied as an educational tool to enhance training in recognizing endoscopic features of *H. pylori* gastritis. Nonetheless, this study was conducted retrospectively at a single institution, which may limit the generalizability of the results. Therefore, further multicenter, prospective studies with larger patient cohorts are warranted to validate these findings and to investigate the integration of this system into routine clinical practice. Further validation should also confirm generalizability across different expertise levels by incorporating a more balanced reader panel. Future studies should also consider evaluating DL model performance by specific gastric regions or developing models that diagnose *H. pylori* using comprehensive video coverage of all gastric regions.

KEY MESSAGE

1. A deep learning model classified *H. pylori* infection on endoscopic images with 97.0% patient-level accuracy and 100% sensitivity.
2. Deep learning assistance improved the mean diagnostic accuracy of 14 endoscopists from 72.4% to 83.2%, with the two experts achieving 97.0% accuracy.
3. Deep learning–assisted endoscopy offers the potential to support the diagnosis of *H. pylori* infection during routine gastroscopy.

REFERENCES

1. Kamboj AK, Cotter TG, Oxentenko AS. *Helicobacter pylori*: the past, present, and future in management. Mayo Clin Proc 2017;92:599-604.
2. Wang YK, Kuo FC, Liu CJ, et al. Diagnosis of *Helicobacter pylori* infection: current options and developments. World J Gastroenterol 2015;21:11221-11235.
3. Crowe SE. *Helicobacter pylori* Infection. N Engl J Med 2019;380:1158-1165.
4. Gisbert JP, de la Morena F, Abraira V. Accuracy of monoclonal stool antigen test for the diagnosis of *H. pylori* infection: a systematic review and meta-analysis. Am J Gastroenterol 2006;101:1921-1930.
5. Attumi TA, Graham DY. Follow-up testing after treatment of *Helicobacter pylori* infections: cautions, caveats, and recommendations. Clin Gastroenterol Hepatol 2011;9:373-375.
6. Malfertheiner P, Megraud F, O'Morain CA, et al. Management of *Helicobacter pylori* infection-the Maastricht V/Florence Consensus Report. Gut 2017;66:6-30.
7. Ho B, Marshall BJ. Accurate diagnosis of *Helicobacter pylori*. Serologic testing. Gastroenterol Clin North Am 2000;29:853-862.
8. Moon SW, Kim TH, Kim HS, et al. United rapid urease test is superior than separate test in detecting *Helicobacter pylori* at the gastric antrum and body specimens. Clin Endosc 2012;45:392-396.
9. Lan HC, Chen TS, Li AF, Chang FY, Lin HC. Additional corpus

biopsy enhances the detection of *Helicobacter pylori* infection in a background of gastritis with atrophy. BMC Gastroenterol 2012;12:182.
10. Seo J, Ahn JY. Endoscopic features according to *Helicobacter pylori* infection status. Korean J Med 2023;98:117-124.
11. Mohan BP, Khan SR, Kassab LL, et al. Convolutional neural networks in the computer-aided diagnosis of *Helicobacter pylori* infection and non-causal comparison to physician endoscopists: a systematic review with meta-analysis. Ann Gastroenterol 2021;34:20-25.
12. Bang CS, Lee JJ, Baik GH. Artificial intelligence for the prediction of *Helicobacter pylori* infection in endoscopic images: systematic review and meta-analysis of diagnostic test accuracy. J Med Internet Res 2020;22:e21983.
13. Kang D, Lee K, Kim J. Diagnostic usefulness of deep learning methods for *Helicobacter pylori* infection using esophagogastroduodenoscopy images. JGH Open 2023;7:875-883.
14. Yacob YM, Alquran H, Mustafa WA, Alsalatie M, Sakim HAM, Lola MS. *H. pylori* related atrophic gastritis detection using enhanced convolution neural network (CNN) learner. Diagnostics (Basel) 2023;13:336.
15. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21-26; Honolulu, HI. New York (NY): IEEE, 2017:2261-2269.
16. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009 Jun 20-25; Miami, FL. New York (NY): IEEE, 2009:248-255.
17. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22-29; Venice, Italy. New York (NY): IEEE, 2017:618-626.
18. Park SH, Han K, Jang HY, et al. Methods for clinical evaluation of artificial intelligence algorithms for medical diagnosis. Radiology 2023;306:20-31.
19. Mu T, Lu ZM, Wang WW, et al. *Helicobacter pylori* intragastric colonization and migration: endoscopic manifestations and potential mechanisms. World J Gastroenterol 2023;29:4616-4627.
20. Toyoshima O, Nishizawa T. Kyoto classification of gastritis: advances and future perspectives in endoscopic diagnosis of gastritis. World J Gastroenterol 2022;28:6078-6089.
21. Cardos AI, Maghiar A, Zaha DC, et al. Evolution of diagnostic methods for *Helicobacter pylori* infections: from traditional tests to high technology, advanced sensitivity and discrimination tools. Diagnostics (Basel) 2022;12:508.
22. Kishikawa H, Ojiro K, Nakamura K, et al. Previous *Helicobacter pylori* infection-induced atrophic gastritis: a distinct disease entity in an understudied population without a history of eradication. *Helicobacter* 2020;25:e12669.
23. Kim GH. Endoscopic findings of Kyoto classification of gastritis. Korean J *Helicobacter* Up Gastrointest Res 2019;19:88-93.

Received : July 14, 2024
Revised : January 20, 2026
Accepted : June 16, 2026

Correspondence to
Do Hoon Kim, M.D., Ph.D.
Department of Gastroenterology, University of Ulsan College of Medicine, Asan Medical Center, 88, Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Korea
Tel: +82-2-3010-3197, Fax: +82-2-3010-6517
E-mail: dohoon.md@gmail.com
https://orcid.org/0000-0002-4250-4683

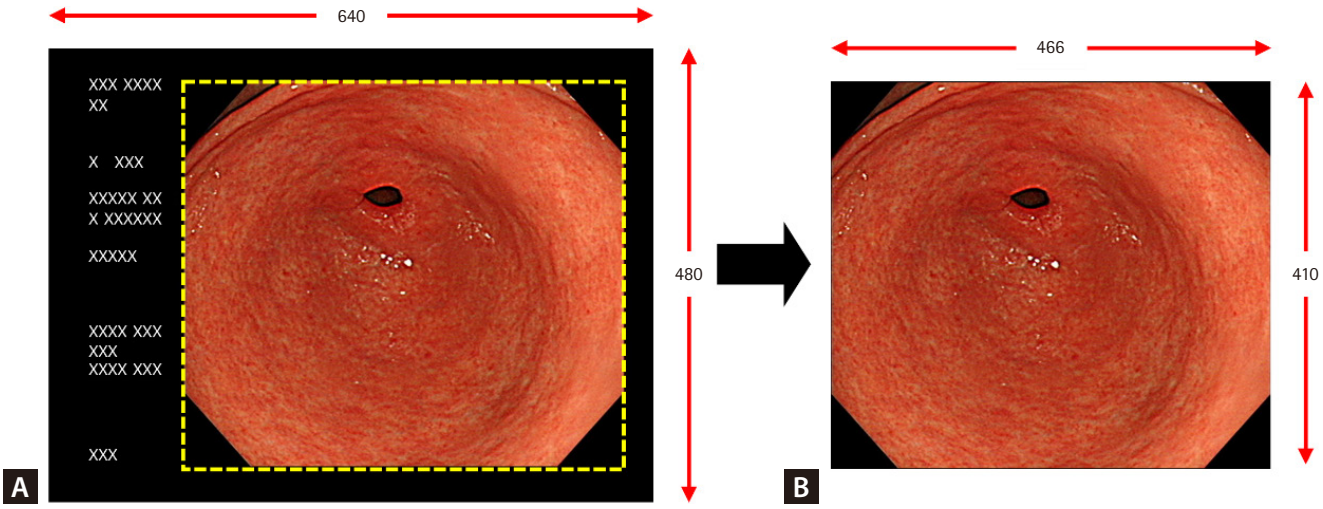
Correspondence to
Namkug Kim, Ph.D.
Department of Convergence Medicine, University of Ulsan College of Medicine, Asan Medical Center, 88, Olympic-ro 43-gil Songpa-gu, Seoul 05505, Korea
Tel: +82-2-3010-6573, Fax: +82-2-476-4719
E-mail: namkugkim@gmail.com
https://orcid.org/0000-0002-3438-2217

CRedit authorship contributions
Jun-young Seo: investigation, data curation, formal analysis, validation, writing - original draft, writing - review & editing; Jiseon Kang: investigation, data curation, formal analysis, validation, software, writing - original draft, writing - review & editing; Do Hoon Kim: conceptualization, formal analysis, writing - review & editing, supervision; Namkug Kim: conceptualization, writing - review & editing, supervision

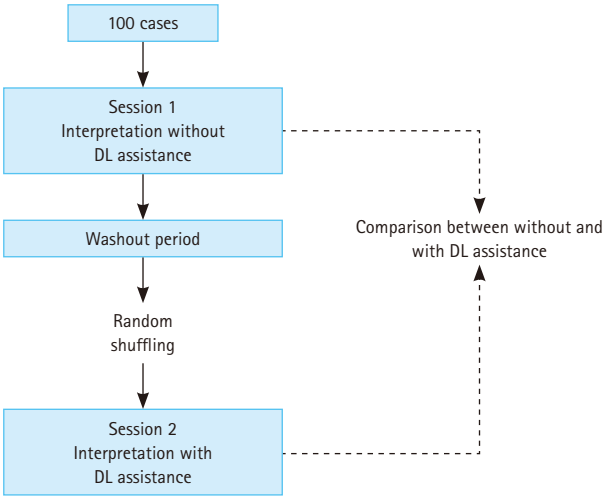
Conflicts of interest
The authors disclose no conflicts.

Funding
None

https://doi.org/10.3904/kjim.2024.245



Supplementary Figure 1. An illustration of preprocessing endoscopic images. (A) The original images stored in PACS contain text related to user information. (B) After preprocessing to extract the area corresponding to the endoscope scene, the image resolution is 466 × 410. PACS, Picture Archiving and Communication System.



Supplementary Figure 2. Study design comparing the impact of using DL algorithms as additional diagnostic tools. DL, deep-learning.

https://doi.org/10.3904/kjim.2024.245

Supplementary Table 1. Data partitions

Dataset		Patients		Images	
		H. pylori-negative	*H. pylori*-positive	*H. pylori*-negative	*H. pylori*-positive
ClearScope-200 (n = 200 patients)	Training	80	80	734	431
	Validation	10	10	91	47
	Test	10	10	84	47
	Total	100	100	909	525
FullScope-100 (n = 100 patients)	Test	50	50	652	653

The ClearScope-200 dataset (200 patients; 1,434 images) is for deep learning model development and divided into training, validation, and test subsets. The FullScope-100 dataset (100 patients; 1,305 images) serves as a test set to compare the performance of endoscopists with and without support of deep learning.
H. pylori, *Helicobacter pylori*.

Supplementary Table 2. Clinical characteristics of patients

Characteristic	ClearScope-200 (n = 200 patients)		FullScope-100 (n = 100 patients)		*p* value
	Current infection (n = 100)	Non-infection (n = 100)	Current infection (n = 50)	Non-infection (n = 50)	
Sex					> 0.999
Male	55	48	28	29	
Female	45	52	22	21	
Age (yr)	55.0 (49.0–62.0)[a)]	59.0 (53.0–66.0)[a)]	54.5 (46.0–63.0)[a)]	52.5 (49.0–57.0)[a)]	0.681
Serum *H. pylori* IgG	> 8.00	0.00	> 8.00	0.00	< 0.001

H. pylori, Helicobacter pylori.
[a)]Data are shown as medians with ranges in parentheses.

https://doi.org/10.3904/kjim.2024.245