




Development and multi-institutional validation of an artificial intelligence-based diagnostic system for gastric biopsy

Hiroyuki Abe^{1,2}  | Yusuke Kurose^{3,4} | Shusuke Takahama⁵ | Ayako Kume¹ | Shu Nishida¹ | Miyako Fukasawa¹ | Yoichi Yasunaga¹ | Tetsuo Ushiku¹ | Youichiro Ninomiya⁶ | Akihiko Yoshizawa^{2,7} | Kohei Murao⁶  | Shin'ichi Sato⁶ | Masaru Kitsuregawa^{6,8} | Tatsuya Harada^{3,4,6} | Masanobu Kitagawa^{2,9} | Masashi Fukayama^{1,2,10}  | Japan Pathology AI Diagnostics/National Institute of Informatics (JP-AID/NII) Study Group for Gastric Biopsy Pathology

¹Department of Pathology, Graduate School of Medicine, the University of Tokyo, Tokyo, Japan

²Japanese Society of Pathology, Tokyo, Japan

³Research Center for Advanced Science and Technology, the University of Tokyo, Tokyo, Japan

⁴Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan

⁵Graduate School of Information Science and Technology, the University of Tokyo, Tokyo, Japan

⁶Research Center for Medical Bigdata, National Institute of Informatics, Tokyo, Japan

⁷Department of Diagnostic Pathology, Graduate School of Medicine, Kyoto University, Kyoto, Japan

⁸Institute of Industrial Science, the University of Tokyo, Tokyo, Japan

⁹Department of Comprehensive Pathology, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, Tokyo, Japan

¹⁰Asahi TelePathology Center, Asahi General Hospital, Chiba, Japan

Correspondence

Masashi Fukayama, Asahi TelePathology Center, Asahi General Hospital, I-1326, Asahi, Chiba, 289-2511, Japan.
Email: mfukayama-ky@umin.org

Funding information

Japan Agency for Medical Research and Development, Grant/Award Number: JP161k1010022, JP181k1010027, JP181k1010028 and JP191k1010036

Abstract

To overcome the increasing burden on pathologists in diagnosing gastric biopsies, we developed an artificial intelligence-based system for the pathological diagnosis of gastric biopsies (AI-G), which is expected to work well in daily clinical practice in multiple institutes. The multistage semantic segmentation for pathology (MSP) method utilizes the distribution of feature values extracted from patches of whole-slide images (WSI) like pathologists' "low-power view" information of microscopy. The training dataset included WSIs of 4511 gastric biopsy tissues from 984 patients. In tissue-level validation, MSP AI-G showed better accuracy (91.0%) than that of conventional patch-based AI-G (PB AI-G) (89.8%). Importantly, MSP AI-G unanimously achieved higher accuracy rates (0.946 ± 0.023) than PB AI-G (0.861 ± 0.078) in tissue-level analysis,

Abbreviations: AI, artificial intelligence; AI-G, artificial intelligence for gastric biopsy; AUC, area under curve; DCNN, deep convolutional neural network; DegA, degree of anomaly; JP-AID, Japan Pathology AI Diagnostics Project; MSP, multistage semantic segmentation for pathology; NII, National Institute of Informatics; PB, patch-based; ROC, receiver operating characteristic; UTH, University of Tokyo Hospital; WSI, whole-slide image.

Hiroyuki Abe and Yusuke Kurose contributed equally to this work.

A list of study group authors and their affiliations are provided in Supporting Information (Document S2).

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2022 The Authors. *Cancer Science* published by John Wiley & Sons Australia, Ltd on behalf of Japanese Cancer Association.

when applied to the cohorts of 10 different institutes (3450 samples of 1772 patients in all institutes, 198–555 samples of 143–206 patients in each institute). MSP AI-G had high diagnostic accuracy and robustness in multi-institutions. When pathologists selectively review specimens in which pathologist's diagnosis and AI prediction are discordant, the requirement of a secondary review process is significantly less compared with reviewing all specimens by another pathologist.

KEYWORDS

artificial intelligence, deep learning, diagnosis, gastric biopsy, pathology

1 | INTRODUCTION

Artificial intelligence has played a crucial part in numerous fields of human research, including medicine. The DCNN is the major driver of this marked development,¹ especially in the field of image analysis, such as in object recognition. A WSI is a digital image of a histologic section on glass slides that is viewed at high magnification through a microscopic lens of a WSI scanner. It has been used for consultation, research, and education related to pathology. However, its daily use is still in its infancy due to the large image size of each WSI (several hundred mega to giga bases) and scanning time (1–3 min per slide).² AI application to WSI is successful not only in the research of cancer pathology³ but in the practice of pathology, for example, to detect cancer in prostate biopsies or sentinel lymph nodes in breast cancer.^{4–6}

Gastric cancer is a common cause of cancer-related death, especially in East Asia,⁷ and gastric biopsy specimens are commonly analyzed in pathology laboratories. In Japan, based on National Database of Health Insurance Claim Information, 4.0 million endoscopic biopsies are performed in 2016, and approximately half were estimated to be gastric biopsies. Furthermore, endoscopic examination is now part of the primary screening program for gastric cancer in Japan and South Korea, which increases the need for second-expert endoscopy and gastric biopsy.^{8–11} Ideally, each pathological diagnosis is reviewed by another expert pathologist (secondary prospective review) to prevent human error in missing or overdiagnosing cancer, which might otherwise result in serious harm to patients. However, the significant error rates when detected by a second pathologist range from 0.1% to more than 10%, depending on the review method and the lesion type. Random or targeted review is one solution, as recommended by the College of American Pathologists.¹² However, pathologists are responsible for all pathological diagnoses; therefore, there is requirement for an alternate solution, such as the application of the AI-review system of WSIs for reducing pathologists' burden of second review (Figure 1A). After each gastric biopsy specimen is diagnosed by a pathologist, all WSIs are checked by an AI system. In case of differences in diagnosis between pathologists and the AI system, final diagnoses are made by pathologists during the second review. Pathologists can refer to the abnormal-marked images by the AI system that they previously

considered as nonneoplastic. Pathologists can also confirm their abnormal or neoplastic diagnosis using the results of the AI system following nonneoplastic prediction by the AI system. To achieve this endpoint and to further improve the system, we tested the newly developed the "MSP" method.¹³ It utilizes the distribution of feature values extracted from patches such as "low-power view" information of WSI,¹³ whereas the conventional "PB" method utilizes only the features on a single patch (like "high-power view" information). The MSP method enables the application of the semantic segmentation to WSI at a practical level of graphics processing unit (GPU) ability.

The feasibility of the system is essential for multi-institutional use. The performance is influenced at least by two factors: (1) Technical factors of slide preparation. The color tone of the H&E stain considerably differs among different laboratories and WSI scanners.¹⁴ (2) Interobserver differences among pathologists. It is not uncommon in daily clinical practice of pathology that a definite distinction is difficult to determine between neoplastic and reactive changes in gastric biopsy specimens. Therefore, there is inconsistent border classification of normal cells and carcinomas even by gastrointestinal pathology specialists; this produces a gray zone diagnosis such as "indefinite for malignancy." Some differences are inevitable among pathologists at different institutes for "indefinite for malignancy" diagnosis; therefore, it is necessary to make fine adjustments in the AI system at each institution.

In the present study, we developed an autonomous AI assessment system for diagnosing gastric biopsy (AI-G) with a new developed MSP algorithm in addition to conventional PB one (PB AI-G).¹³ Furthermore, to overcome the multi-institutional problem as described above, we organized a nationwide group to collect annotated WSI, named JP-AID. The members from the 10 institutions did not share their training program for pathologists, and the group was suitable for evaluating the usefulness of AI-G. The results showed that the MSP AI-G system is feasible in daily pathological practice, including the WSI-based telepathology network in the regions where the number of pathologists is insufficient.

2 | MATERIALS AND METHODS

Figure 1B shows the flowchart of the study.

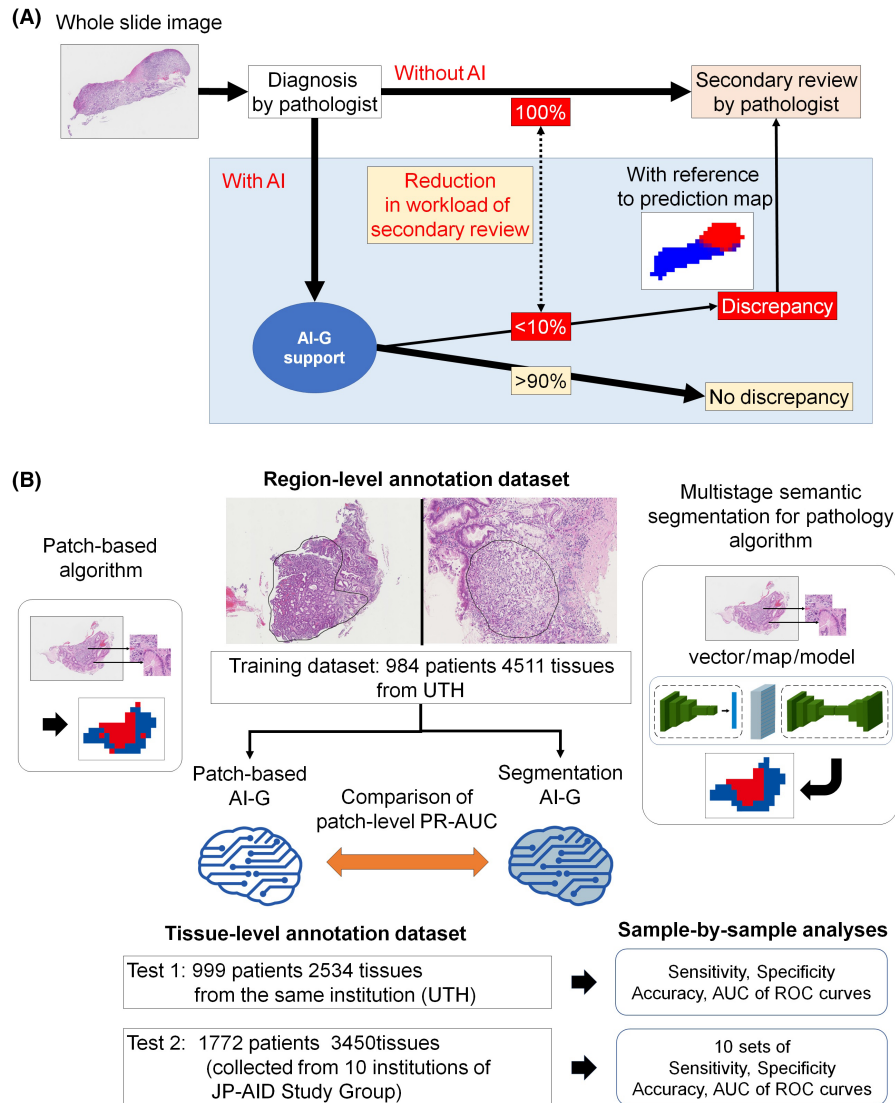


FIGURE 1 Proposed workflow for pathological practice incorporating AI; flowchart showing the construction and validation of the AI system for pathological diagnosis of gastric biopsies. (A) Proposed workflow for pathological practice incorporating AI. Following diagnosis of each gastric biopsy specimen by a pathologist, all WSIs are checked by the AI system. In case of discrepancy, the pathologist reviews the specimen for the final diagnosis. The rate of discrepancy should be less than 10% for AI to be acceptable in daily pathological practice. In this workflow, pathologists can take responsibility for all specimens. (B) Two artificial intelligence-based systems using the conventional patch-based and multistage semantic segmentation for pathology (PB AI-G and MSP AI-G, respectively) similarly trained with region-level annotated datasets (uppermost figures). Both systems are tested with two datasets. One dataset is from the same institution as that of the training dataset (Test 1), whereas the other dataset is collected from multi-institutions of the JP-AID Study Group for validation (Test 2). Sensitivity, specificity, accuracy, and AUC of ROC curves are compared in both tests. AI-G, artificial intelligence system for the pathological diagnosis of gastric biopsies; AUC of ROC curves, area under the curve of receiver operating characteristic curves; JP-AID, Japan Pathology AI Diagnostics Study Group; UTH, University of Tokyo Hospital.

2.1 | Dataset

For the training dataset, the WSIs of 4605 tissue samples were annotated at the region level. The details of the training dataset and annotations are described in Document S1. To confirm the diagnosis, all delineated or encircled areas were annotated according to the Group Classification defined in the Japanese Classification of Gastric Carcinoma (3rd English edition)¹⁵ as follows: Group 1 (G1): normal or nonneoplastic lesion; Group 2 (G2): diagnosis of

neoplastic or nonneoplastic lesion is difficult (indefinite for neoplasia); Group 3 (G3): adenoma; Group 4 (G4): neoplastic lesion that is suspected to be carcinoma; and Group 5 (G5): carcinoma. The number of tissues was 2579 (G1), 48 (G2), 17 (G3), 29 (G4), and 1932 (G5). Only G1 and G5 (4511 tissues) were used for training, and detailed information on training dataset is presented in Table S1.

Both PB- and MSP-AIG were trained using the region-level dataset and applied to the validation datasets. Malignant neoplasms

other than primary gastric carcinoma such as lymphoma and metastatic carcinoma were excluded and separately analyzed.

The validation datasets were the tissue-level annotated dataset, and sample-by-sample analyses were performed. For Test 1 shown in Figure 1, H&E slides were from 999 consecutive patients undergoing gastric biopsy; finally, 2534 tissue samples were obtained at UTH in 2016 (Table S2). For the validation test of samples from multi-institutions (Test 2 in Figure 1B) (Table S3), 3450 tissues from 1772 patients were collected from 10 institutions (median, 334 tissues and 188.5 patients; range, 198–555 tissues and 143–206 patients per institution). As shown in Table S3, H&E staining methods and case selection (consecutive or selected for specific cases) varied among institutions. Different WSI scanners were used among the institutions: eight institutions used NanoZoomer (Hamamatsu Photonics), two used Scanscope (Leica, Wetzlar, Germany), and the other one used Ventana DP (Roche, Basel, Switzerland). OpenSlide, which is the library to provide an interface for reading a various WSI format, enabled us to directly analyze WSIs produced by different scanners.

The study was approved by the institutional review board of the Faculty of Medicine of the University of Tokyo (review number 11603) and the Japanese Society of Pathology (review number 003-2018). The study was performed in accordance with the Declaration of Helsinki.

2.2 | Machine learning

We developed two AI assessment system; “PB” and “multistage semantic segmentation for pathology WSI (MSP)”. The DCNN was used for feature extraction and for the classifier. GoogLeNet (Google Inc., Mountain View, CA, USA)¹⁶ was adopted as the DCNN structure, because GoogLeNet is a light network with less parameters compared with other neural networks used for image classification, which is an advantage to popularize the developed model all over the world with less expensive GPU. During training and inference, we used a single GPU (NVIDIA Tesla V100 32GB) in NVIDIA DGX-1. However, we confirmed that our AI system training could run on a single GPU with 16GB memory. Details of machine learning is described in Document S1.

To confirm the technical validity of the machine learning, we evaluated the performance of PB and MSP AI-G by five-fold cross-validation; five groups of samples were made (collecting the samples from the same patient into one), using four groups for training and one group for testing, respectively. In the training group, 5% of the training data were assigned to validation data to optimize the deep learning. We compared the pathologist’s annotation with our model’s prediction in each patch and calculated the precision recall area under the curve (PR-AUC). Experiments were repeated three times for each combination with the training set and test set. The output of both models was displayed as a heat map that showed the degree of anomaly (DegA; values from 0–1) as a color gradation

of blue and red. The average number of patches extracted from a single tissue sample was 169.4 ± 99.5 for G1 and 58.7 ± 61.8 for G5. The patch-level concordance between the pathologist’s annotation and the prediction of G1 and G5 demonstrated that PR-AUC were 0.959, on average, for conventional PB AI-G and 0.990 for MSP AI-G ($p < 0.001$) (Table S4).

We evaluated the feasibility of AI-G using WSIs for the tissue-level annotated dataset by sample-by-sample analysis (Tests 1 and 2 in Figure 1). The prediction ability of AI-G was determined as follows: First, the DegA of each patch (256×256 pixels) was calculated using AI-G. The value of DegA ranges from 0 to 1, and lower DegA implies a higher probability of G1. When DegA was more than the cutoff value in at least one patch, the tissue was classified as “G5” or “G2–5.” When all the patches in the tissue showed DegA below the cutoff value, the tissue was classified as “G1.” The prediction ability of both models was compared with the original pathological diagnosis (G1 and G5, and G1 and G2, G3, G4, G5) (Figure S1). We then calculated the following indices: sensitivity for G5, sensitivity for G2–G5, specificity for G1, and concordance rate (accuracy) (Figure S1). Next, we created a ROC curve and calculated the AUC based on the results with various cutoff DegA values (0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, and 0.99). The optimal cutoff DegA values, in which both sensitivity and specificity are near 1, was also calculated in each institution. Individual sensitivities for G2, G3, and G4 were also calculated using the optimal cutoff DegA value of G1/G2–5. These evaluations were performed for both PB and MSP AI-G. In multi-institutional analyses, the original diagnosis of the individual institution was adopted as a ground truth because the discrepancy between pathologists could also affect the accuracy of AI-G.

2.3 | Color normalization of WSIs and its effect on the assessment of AI-G

Color normalization of WSIs was performed to assess the effect of color variation on optimal cutoff values in each institution. Specifically, the distribution of red, green, and blue values in each institution was calculated and adjusted to that of UTH, the institution where the training datasets were obtained. After color normalization, ROC curves were constructed and optimal cutoff values were calculated in each institution.

2.4 | Statistical analyses

We used a paired or unpaired *t*-test for continuous variables. For comparisons of a pathologist’s diagnosis and the prediction of our model, sensitivity, specificity, accuracy, and AUC of the ROC curve were calculated. All statistical analyses were performed using EZR software (Saitama Medical Center, Jichi Medical University, Saitama, Japan), which is a graphical user interface for R (The R Foundation for Statistical Computing, Vienna, Austria). More precisely, EZR is a

modified version of R commander designed to add statistical functions frequently used in biostatistics.¹⁷

3 | RESULTS

3.1 | Performance of PB AI-G and MSP AI-G

Preliminary study demonstrated that MSP AI-G was superior to PB AI-G even in patch-level comparison, as described in Materials and Methods (Table S4). Representative images of the original WSI and the AI prediction heat map of both AI-Gs are shown in Figure 2 and Figure S2. The heat map was displayed so that the degree of anomaly (DegA; values from 0–1) was presented as a color gradation from blue (DegA = 0) to red (DegA = 1). Both AI-Gs detected gastric carcinoma of two major types, namely, well differentiated tubular adenocarcinoma, as shown in Figure 2A,E, and poorly cohesive adenocarcinoma, as shown in Figure 2F,J. In the former case, MSP AI-G frequently detected a larger area than PB AI-G because of the inclusion of cancer glands at the border to normal glands. Furthermore, MSP AI-G excluded scattered false-positive patches of PB AI-G in the nonneoplastic mucosa (Figure 2C,D). In the poorly differentiated adenocarcinoma, MSP AI-G also detected cancer cells at the border in addition to the area detected by PB AI-G (Figure 2H,I).

3.2 | Comparison of prediction ability between PB and MSP AI-Gs (Test 1)

The prediction abilities of both AI-Gs were further evaluated using WSI images of different datasets of gastric biopsy samples (2534 tissue samples [2107 G1, 123 G2-4, and 304 G5] of 999 patients) with the pathologist's tissue-level annotation from the same institute (UTH) (Table S2).

All the data in the analyses below were results on a sample-by-sample basis. To compare the prediction abilities of both AI-Gs, the sensitivity and specificity (Figure S1) were obtained with different DegAs to generate receiver operating characteristic (ROC) curve (Figure 3A–D). Analyses were performed in two ways: to compare the classification of carcinoma (G1/G5) or abnormal (G1/G2–5). In the former, only G1 (2107 samples) and G5 samples (304 samples) were used. In the latter, which was similar to real daily clinical practice, G2–4 samples (123 samples) were grouped and all 2534 samples were used for evaluation. AUCs of MSP AI-G and PB AI-G were almost equal in both analyses (Figure 3A, B for G1/G5 and Figure 3C, D for G1/G2–5), but the configuration of AUC was more flexible in MSP AI-G, as shown by the changes in optimal DegA. The results showed a difference of increased specificity and accuracy by MSP AI-G (91.3% and 91.0%) compared with those by PB AI-G (89.2% and 89.8%), which affected the number of samples necessary for

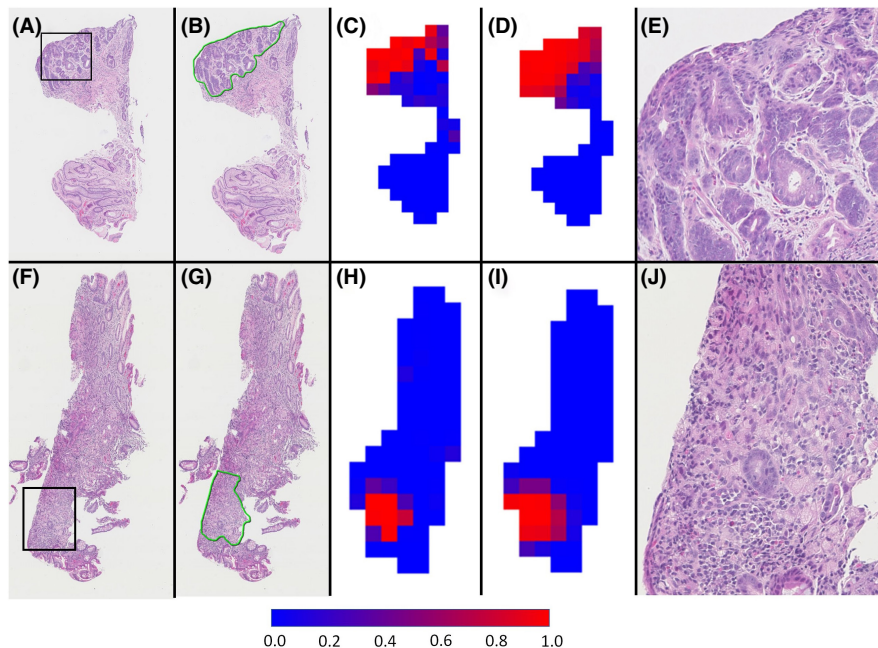


FIGURE 2 Original images of gastric cancer biopsy and prediction heat maps prepared using PB AI-G and MSP AI-G. The probability of adenocarcinoma (degree of anomaly) is shown as a color gradation from deep blue, indicating low probability, to deep red, indicating high probability. Two panels are examples of a well differentiated adenocarcinoma (A–E) and a poorly cohesive adenocarcinoma with signet ring cells (F–J). The panels are arrayed as original whole-slide images (A, F) annotated images by pathologists (ground truth distribution delineated in green) (B, G), prediction heat maps prepared using PB AI-G (C, H) and MSP AI-G (D, I), and high-power views of the original whole-slide images (E, J; corresponding to the black squares in A and F). Both AI-G detected cancer regions, almost corresponding to the annotated regions; their borders were delineated in (B) and (G). However, patches with intermediate color between blue and red markedly decreased in MSP AI-G (D, I) compared with PB AI-G (C, H). Heatmaps of MSP AI-G corresponded more to the annotated cancer region. AI-G, artificial intelligence system for the pathological diagnosis of gastric biopsies; MSP, multistage semantic segmentation for pathology; PB, patch-based.

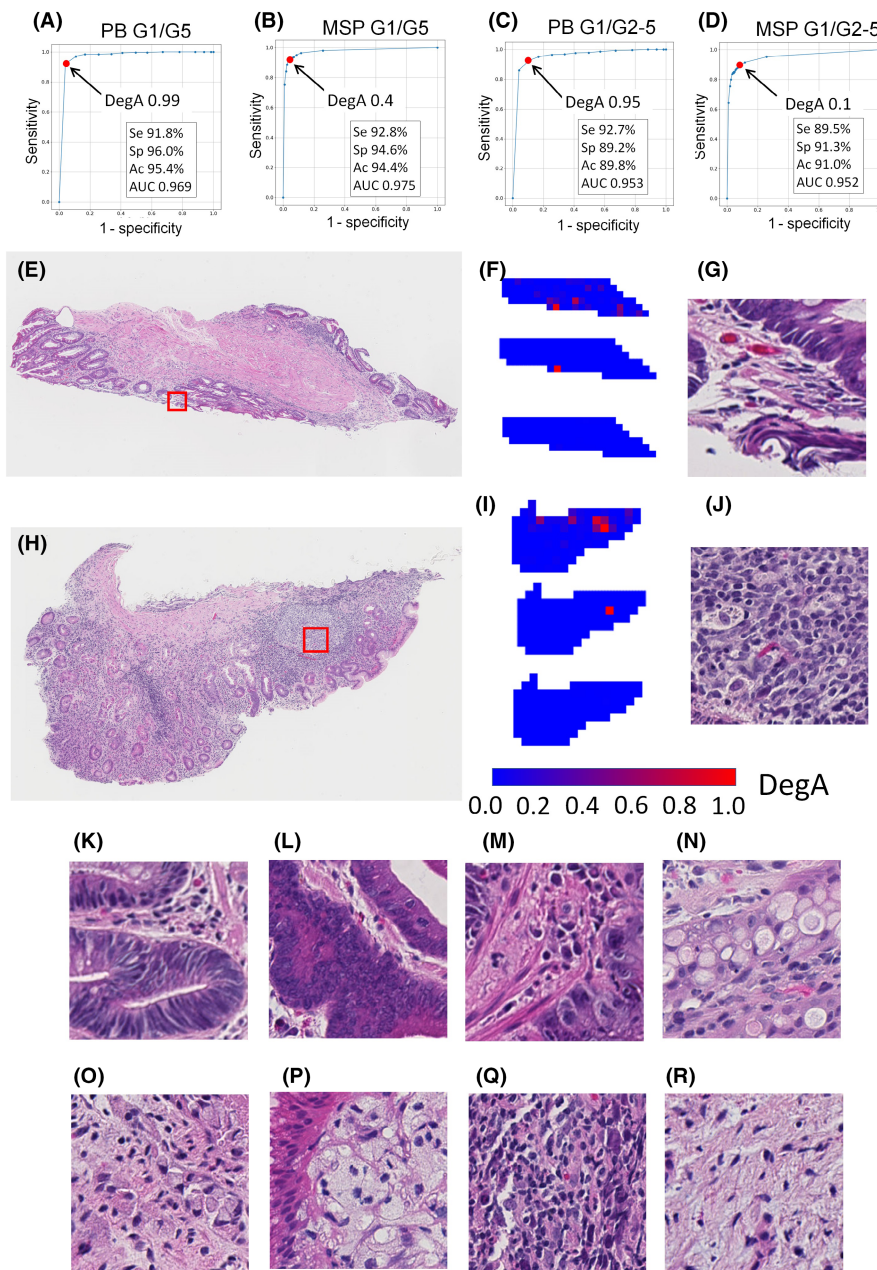


FIGURE 3 PB AI-G and MSP AI-G applied to validation datasets from the same institute (Test 1). Receiver operating characteristic (ROC) curves were constructed for G1/G5 classification by PB AI-G (A) and MSP AI-G (B), and for G1/G2-5 classification by PB AI-G (C) and MSP AI-G (D). The red point indicates the nearest cutoff value that makes both sensitivity and specificity close to 1. Panels (E-G and H-J) are examples of G1 diagnosed by pathologists, neoplasm predicted by PB AI-G, and normal predicted by MSP AI-G with the cutoff value for G1/G2-5. The panels are arrayed as original whole-slide images, heat maps, and discordant patches. There are three heat maps (F, I). The top shows DegA predicted by PB AI-G as a color gradation (blue to red), the middle shows patches where DegA predicted by PB AI-G was more than 0.95 as red, and the bottom shows DegA predicted by MSP AI-G as a color gradation (blue to red). Each red square in (E) or (H) points out the location of the false-positive patch in PB AI-G (G) or (J), respectively. Other patch images (K-R) are examples of G1 diagnosed by pathologists, neoplasia predicted by PB AI-G, and normal predicted by MSP AI-G with the cutoff value of G1/G2-5; glandular cells with reactive enlarged nuclei (K-M), degenerative glandular cells (N, O), and various types of stromal cells (P-R). Ac, accuracy; AI-G, artificial intelligence system for gastric biopsy; AUC, area under the curve; MSP, multistage semantic segmentation for pathology; PB, patch-based; ROC, receiver operating characteristic; Se, sensitivity; Sp, specificity.

secondary review. In total, 94 G1 tissues were wrongly assigned as G5 by PB AI-G but correctly assigned as G1 by MSP AI-G, whereas 49 G1 tissues were wrongly assigned as G5 by MSP AI-G but correctly assigned as G1 by PB AI-G. As a result, the number of G1

tissues wrongly assigned as G5 decreased by 45 from 228 to 183 by MSP AI-G.

To evaluate the characteristics of MSP AI-G performance, we chose patch images of G1 evaluated by both pathologists and MSP

AI-G, which were assigned as neoplasia by PB AI-G (106 patches from 94 tissues) (Figure 3E-R). Forty-nine (46%) patches were found in the peripheral location (Figure 3E-G), and 57 (54%) patches were found in the central location (Figure 3H-J). In total, 83 patches (78%) included glands with reactive nucleus enlargement (Figure 3G, K-O), and 48 patches (45%) contained considerable amounts of the stromal component, such as inflammatory cells and xanthoma cells (Figure 3J, P-R). These glands and stromal cells might be consistent with atypia within a limited frame of visual fields.

3.3 | Multi-institutional validation of PB and MSP AI-Gs

To further validate the feasibility of MSP AI-G in practical application, WSIs of gastric biopsy were collected from 10 institutions of the JP-AID Study Group other than UTH. The number of tissue samples was 3450 in total (range 198–555 samples, consisting of G1, 2514; G2–G4, 350; and G5, 586) (Table S3). The pathologist's diagnosis and AI-G prediction of each sample were compared. ROC curves of each institution are presented in Figure 4A–D, and sensitivities, specificities, accuracies, AUCs, and optimal thresholds of DegA are summarized in Table S5. MSP and PB AI-Gs are compared in bar graphs with the mean data of mean of 10 institutes and UTH (Figure 4E–H). When the analysis was restricted to G1 and G5, the AUC of MSP AI-G was higher than that of conventional PB AI-G in all institutions ($p < 0.001$, paired *t*-test) (Figure 4E). As for the prediction ability of G1/G2–5, the AUC of MSP AI-G was higher than that of conventional PB AI-G in nine of the 10 institutions ($p < 0.001$, paired *t*-test) (Figure 4F). Accuracy (Figure S1) was also significantly higher in MSP AI-G than in PB AI-G ($p = 0.002$ (paired *t*-test) for G1/G5, and $p = 0.002$ (paired *t*-test) for G1/G2–5, respectively) (Figure 4G,H). The reason for the high performance of MSP AI-G compared with that of PB AI-G was derived from a marked decrease in the number of false-positive samples in MSP (Figure S3).

To evaluate the prediction of G2–4 by both AI-Gs in detail, the results of G2, G3, and G4 in UTH and the other 10 institutions were analyzed. The results from 10 institutions were combined because they varied considerably in number (Figure S4). When the optimal cutoff value of G1/G2–5 (Table S5) was adopted, the sensitivity of G2–4 was not different in both AI-Gs, but the percentage of AI prediction was lowest in G2 in MSP AI-G.

3.4 | Potential factors influencing multi-institutional differences

There were several potential factors influencing the differences among multiple institutions. Consecutive cases were analyzed in eight of the 10 institutions, but the selected samples, which included more samples of G2–5, were analyzed in two institutions (D and I) (Table S3). The same WSI scanners were used in seven institutions

with that in UTH, but they were different in the institutes C, D, and J. There appeared no differences in the AUC of the ROC curve (Figure 4E, F) and accuracy (Figure 4G, H) in these institutes compared with those of others.

To estimate the effect of color variation on the performance of both AI-Gs, the distribution of red, green, and blue values in each institution was adjusted to that of UTH (Tables S6 and S7), from which images were used for the development of AI-Gs. An example of color normalization is presented in Figure S5. In G1/G5 and G1/G2–5 testing (Figure S6), color normalization did not affect the AUC of MSP AI-G, when optimization was performed. However, the effect on PB AI-G was considerable in four institutions, but unpredictable.

3.5 | Samples escaping from both PB AI-G and MSP AI-G

In UTH test datasets, nine (3.0%) of 304 G5 samples were classified as G1 by both AI-Gs. Three of them were fundic gland-type adenocarcinoma (oxyntic type adenoma) with minimal nuclear atypia (Figure S7A,B). When samples of neoplasms other than gastric carcinoma or adenoma were tested (Table S8), both AI-G detected diffuse large B-cell lymphoma, gastrointestinal stromal tumor, neuroendocrine tumor, and metastatic melanoma. However, low-grade lymphomas (mucosa-associated lymphatic tissue [MALT] lymphoma, mantle cell lymphoma, and follicular lymphoma) were missed by both PB and MSP AI-Gs (Figure S7C,D).

4 | DISCUSSION

Autonomous AI assessment system for gastric biopsy evaluation (AI-G) has been constructed using the WSI MSP method, which we recently developed to process gigabase-sized WSIs in daily clinical practice of diagnostic pathology with standard GPU abilities. MSP AI-G enables us to integrate global information of the whole-tissue-scale to high-resolution local information, such as the way of human pathologists correcting the impression at high-power view by low-power view. MSP AI-G showed higher patch-level performance than the conventional PB method (PB AI-G) and improved accuracy and specificity in the validation tests of G1/G2–5 classification in gastric biopsy. Furthermore, in the multi-institutional study of the JP-AID Study Group, MSP AI-G achieved higher performance in both the G1/G5 and G1/G2–5 tests in samples collected from other 10 institutions. There are already a few reports on the use of AI in the detection of cancer in gastric biopsy, some of which showed almost equal accuracy to our AI-G; however, they were tested in only two or three institutions including the same institution as that of the training datasets.^{18–21} These results proved the effectiveness of the multistage segmentation approach to image recognition in diagnostic pathology. In future, MSP AI-G is highly recommended for implementation in daily pathological practice.

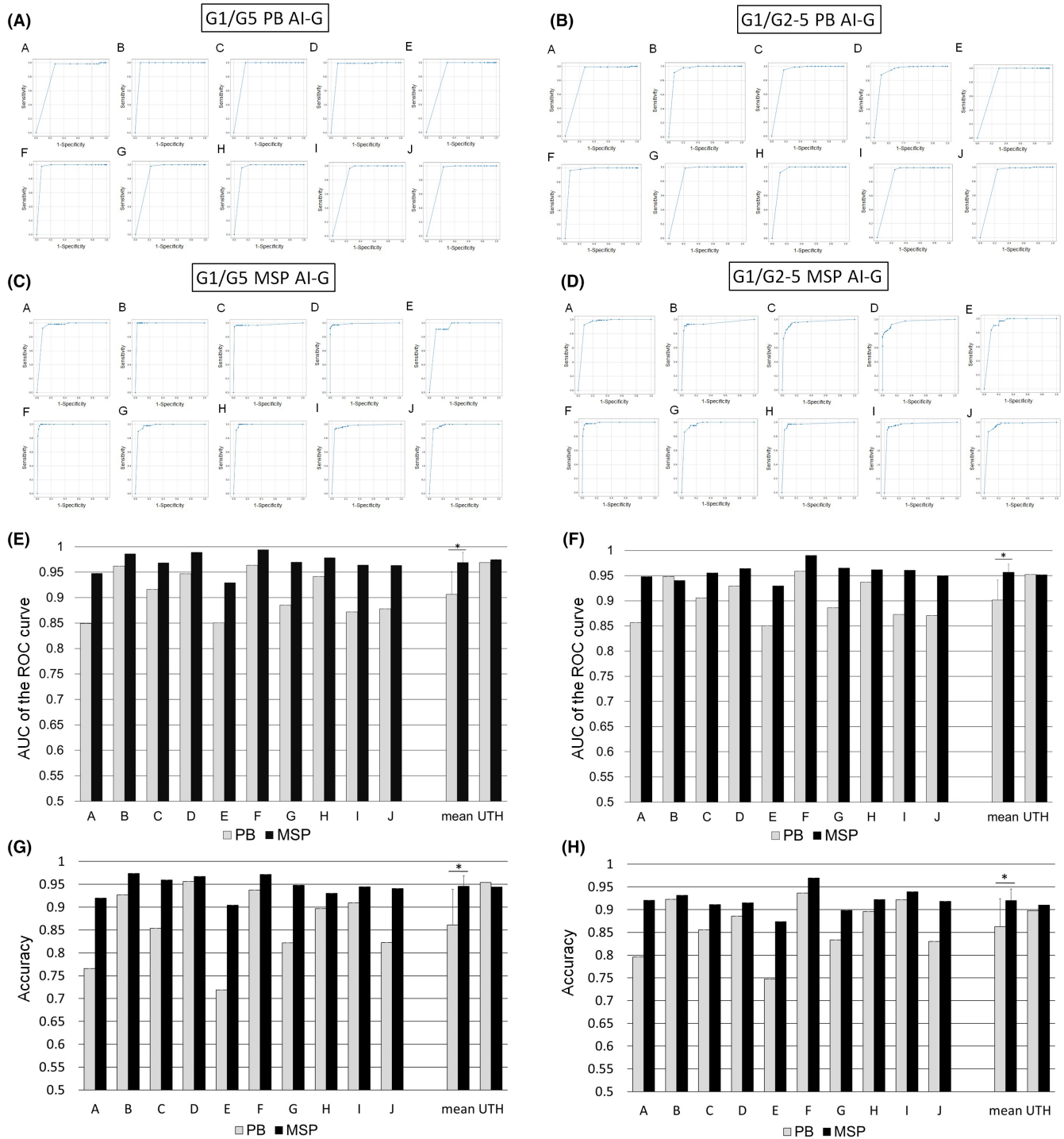


FIGURE 4 PB AI-G and MSP AI-G applied to validation datasets from multiple institutes of the JP-AID Study Group (Test 2). ROC curves were constructed for G1/G5 classification by PB AI-G (A) and MSP AI-G (B) and MSP AI-G (C) and for G1/G2-5 classification by PB AI-G (D). Bar graphs (E, F) show AUC values of ROC curves by PB AI-G and MSP AI-G in 10 institutes for G1/G5 classification and G1/G2-5 classification, respectively. The data from the University of Tokyo Hospital (institute of the training dataset) were also included in the bar graphs. Bar graphs (G, H) showing the accuracies of PB AI-G and MSP AI-G in the 10 institutes are compared for G1/G5 classification and G1/G2-5 classification, respectively; higher performance of MSP AI-G than that of PB AI-G. AUC, area under the curve; MSP, multistage semantic segmentation for pathology; PB, patch-based; ROC, receiver operating characteristic; UTH, the University of Tokyo Hospital. * $p < 0.001$.

There is currently no consensus on how to integrate AI into daily pathological practice. AI prediction is not allowed to substitute for pathologists in any fields unless 100% reliability is gained in general;

currently, pathologists take responsibility for every pathological diagnosis. However, it is acceptable to use AI prediction to check the pathologist's diagnosis. In case of discrepancy between the

pathologist's diagnosis and AI prediction, the diagnosis is re-checked by the pathologist and corrected if necessary (Figure 1A). First, we set our endpoint to extract 10% of the total specimens at most, that is, to achieve more than 90% accuracy of AI-G. The predictive power of the system is a function of sensitivity and specificity, and the incidence of the events. The incidence of gastric cancer in biopsy specimens is reportedly 5%–10% in community hospitals,^{22,23} and the specificity should be higher than 89.5% to accomplish this condition when sensitivity is 95% according to $[(\text{sensitivity} \times \text{incidence}) + (\text{specificity} \times (1 - \text{incidence}))]$. With these considerations, in this study, the accuracy of MSP AI-G was 91.3%, which means that the percentage of WSI images necessary for review by human pathologists is less than 9.0%. Furthermore, the pathologists can refer to the heat map of MSP AI-G at their second review, which creates benefits for them to check the lesions at glance.

The usefulness of the AI system, such as AI-Gs, should be proven by applying it to different datasets of other institutes before it is widely accepted. In the present study, to evaluate robustness to this issue, we conducted a multi-institutional validation study and collected 3450 WSIs from 10 institutes in Japan. Although there were many differences in WSIs of each institute, such as the collection method (consecutive or selective) and the proportion of G2–5 and scanner instruments, MSP AI-G showed higher performance and flexibility in all institutes than conventional AI-G after appropriate optimization. Differences in the color of H&E staining may be a possible factor relating to decreased performance of PB AI-G. Adjusting the color distribution to that of originally developed datasets improved the performance of PB AI-G in half of the institutes, but the effect was barely predictive, while the performance of MSP AI-G was quite stable. Therefore, MSP AI-G is robust for color variation, and color adjustment might not be necessary for MSP AI-G.

In the present study, we considered two situations; G1/G5 classification and G1/G2–5 classification. It depends on the situation for pathologists which classification system is adopted in daily clinical practice. It is possible to restrict secondary review to G1 and G5 samples with the G1/G5 cutoff value. However, it is more practical to submit all of the samples to MSP AI-G for the selection of secondary review.

There are limitations of the present study. Both AI-Gs could not recognize abnormalities in oxyntic gland adenoma/fundic gland adenocarcinoma and low-grade lymphomas, in which nuclear atypia of the epithelial cells or lymphocytes was minimal.^{24,25} However, both types of neoplasms are difficult to diagnose properly by human pathologists without any knowledge of clinical information including endoscopy findings. Integrating this information into the AI-G system might increase its overall performance. Implementation of the AI-G in daily clinical practice of pathological laboratories is another problem because it is necessary to digitize glass slides before applying the AI-G. The telepathology network, which sends WSIs from regional hospitals that do not have pathologists to institutions with pathologists, might be a likely candidate to apply the AI-G. The JP-AID Study Group is now addressing this issue by using AI-G in the

regional telepathology networks for pathological diagnoses, that is WSIs are sent from regional hospitals to a core hospital through the cloud server; pathologists diagnose these WSIs using an AI-assisted review system.

In conclusion, we developed AI-G, an autonomous prediction system for gastric biopsy, which was validated in multi-institutions. The concordance rate for AI-G, especially that of MSP AI-G, was sufficiently satisfactory. The present study showed the possibility of machine learning applications in daily pathological practice. Using AI-G, we can extract samples in which the pathologist's diagnosis and AI prediction are discordant. If pathologists review these extracted samples, efficient and feasible secondary reviews are possible even if a single pathologist is employed at the hospital.

ACKNOWLEDGMENTS

This research was performed as a part of the Japan Pathology AI Diagnostics Project (JP-AID) conducted by Japanese Society of Pathology. This research was supported by AMED under Grant Number JP16lk1010022, JP18lk1010027, JP18lk1010028, and JP19lk1010036. The authors thank Momoe Tsuchida and Maki Sawada for their excellent technical assistance. We would like to thank Editage (www.editage.com) for English language editing.

DISCLOSURES

Shusuke Takahama, Yusuke Kurose, Tatsuya Harada, Hiroyuki Abe, Akihiko Yoshizawa, Masashi Fukayama, and Masanobu Kitagawa have the right to a patent for machine learning algorithm named "multistage semantic segmentation for pathology" (number P2021-56571A). Masashi Fukayama, one of the authors of this article, is an editorial board member of Cancer Science. The other authors declare no conflict of interests.

ETHICS STATEMENT

Approval of the research protocol by an Institutional Review Board: The study was approved by the institutional review board of the Faculty of Medicine of the University of Tokyo (review number 11603) and the Japanese Society of Pathology (review number 003–2018). Informed Consent: N/A. Registry and the Registration No. of the study/trial: N/A. Animal Studies: N/A.

ORCID

Hiroyuki Abe  <https://orcid.org/0000-0003-4513-1007>

Kohei Murao  <https://orcid.org/0000-0002-1556-9406>

Masashi Fukayama  <https://orcid.org/0000-0002-0460-064X>

REFERENCES

1. Miller DD, Brown EW. Artificial intelligence in medical practice: the question to the answer? *Am J Med*. 2018;131:129–133. doi:10.1016/j.amjmed.2017.10.035
2. Salto-Tellez M, Maxwell P, Hamilton P. Artificial intelligence—the third revolution in pathology. *Histopathology*. 2019;74:372–376.
3. Komura D, Kawabe A, Fukuta K, et al. Universal encoding of pan-cancer histology by deep texture representations. *Cell Rep*. 2022;38:110424.

4. Steiner DF, Macdonald R, Liu Y, et al. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *Am J Surg Pathol*. 2018;42:1636-1646.
5. Litjens G, Sánchez CI, Timofeeva N, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci Rep*. 2016;6:1-11.
6. Bejnordi BE, Veta M, Van Diest PJ, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*. 2017;318:2199-2210.
7. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68:394-424.
8. Hamashima C, Okamoto M, Shabana M, Osaki Y, Kishimoto T. Sensitivity of endoscopic screening for gastric cancer by the incidence method. *Int J Cancer*. 2013;133:653-659.
9. Hamashima C. Benefits and harms of endoscopic screening for gastric cancer. *World J Gastroenterol*. 2016;22:6385-6392.
10. Hamashima C, Goto R. Potential capacity of endoscopic screening for gastric cancer in Japan. *Cancer Sci*. 2017;108:101-107.
11. Jun JK, Choi KS, Lee HY, et al. Effectiveness of the Korean National Cancer Screening Program in reducing gastric cancer mortality. *Gastroenterology*. 2017;152:1319-1328.e7.
12. Nakhleh RE, Nosé V, Colasacco C, et al. Interpretive diagnostic error reduction in surgical pathology and cytology: guideline from the college of American pathologists pathology and laboratory quality center and the association of directors of anatomic and surgical pathology. *Arch Pathol Lab Med*. 2016;140:29-40.
13. Takahama S, Kurose Y, Mukuta Y, et al. Multi-stage pathological image classification using semantic segmentation. *Proc IEEE Int Conf Comput Vis*. 2019;2019:10701-10710.
14. Komura D, Ishikawa S. Machine learning approaches for pathologic diagnosis. *Virchows Arch*. 2019;475:131-138.
15. Japanese Gastric Cancer Association. Japanese classification of gastric carcinoma: 3rd English edition. *Gastric Cancer*. 2011;14:101-112.
16. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. *Proc IEEE Conf Comput Vis Pattern Recognit*. 2015;1-9.
17. Kanda Y. Investigation of the freely available easy-to-use software 'EZR' for medical statistics. *Bone Marrow Transplant*. 2013;48:452-458.
18. Yoshida H, Shimazu T, Kiyuna T, et al. Automated histological classification of whole-slide images of gastric biopsy specimens. *Gastric Cancer*. 2018;21:249-257.
19. Cosatto E, Laquerre P-F, Malon C, et al. Automated gastric cancer diagnosis on H&E-stained sections; Itraining a classifier on a large scale with multiple instance machine learning. *Med Imaging 2013 Digit Pathol*. 2013;8676:867605.
20. 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;10:1-11.
21. Song Z, Zou S, Zhou W, et al. Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning. *Nat Commun*. 2020;11:1-9.
22. Kaizaki Y, Hosokawa O, Miyanaga T, et al. Clinicopathological study of cases indefinite for neoplasia, Group 2, for gastric biopsy. *Stomach and Intestine (Tokyo)*. 2012;47:187-195.
23. Onodera M, Nakashima Y, Takahashi S, Ishiguro S. Pathological profiles of biopsied specimens diagnosed as "Group 2" in the pathological laboratory of a single institution. *Stomach and Intestine (Tokyo)*. 2012;47:196-202.
24. Ueyama H, Yao T, Nakashima Y, et al. Gastric adenocarcinoma of fundic gland type (chief cell predominant type): proposal for a new entity of gastric adenocarcinoma. *Am J Surg Pathol*. 2010;34:609-619.
25. Schechter NR, Yahalom J. Low-grade MALT lymphoma of the stomach: a review of treatment options. *Int J Radiat Oncol Biol Phys*. 2000;46:1093-1103.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Abe H, Kurose Y, Takahama S, et al. Development and multi-institutional validation of an artificial intelligence-based diagnostic system for gastric biopsy. *Cancer Sci*. 2022;113:3608-3617. doi: [10.1111/cas.15514](https://doi.org/10.1111/cas.15514)