Gastric Cancer (2018) 21:249–257
https://doi.org/10.1007/s10120-017-0731-8

ORIGINAL ARTICLE

Automated histological classification of whole-slide images of gastric biopsy specimens

Hiroshi Yoshida[1] · Taichi Shimazu[2] · Tomoharu Kiyuna[3] · Atsushi Marugame[4] · Yoshiko Yamashita[3] · Eric Cosatto[5] · Hirokazu Taniguchi[1] · Shigeki Sekine[1,6] · Atsushi Ochiai[1,7]

Received: 20 February 2017 / Accepted: 25 May 2017 / Published online: 2 June 2017

Abstract
Background Automated image analysis has been developed currently in the field of surgical pathology. The aim of the present study was to evaluate the classification accuracy of the e-Pathologist image analysis software.
Methods A total of 3062 gastric biopsy specimens were consecutively obtained and stained. The specimen slides were anonymized and digitized. At least two experienced gastrointestinal pathologists evaluated each slide for pathological diagnosis. We compared the three-tier (positive for carcinoma or suspicion of carcinoma; caution for adenoma or suspicion of a neoplastic lesion; or negative for a neoplastic lesion) or two-tier (negative or non-negative) classification results of human pathologists and of the e-Pathologist.
Results Of 3062 cases, 33.4% showed an abnormal finding. For the three-tier classification, the overall concordance rate was 55.6% (1702/3062). The kappa coefficient was 0.28 (95% CI, 0.26–0.30; fair agreement). For the negative biopsy specimens, the concordance rate was 90.6% (1033/1140), but for the positive biopsy specimens, the concordance rate was less than 50%. For the two-tier classification, the sensitivity, specificity, positive predictive value, and negative predictive value were 89.5% (95% CI, 87.5–91.4%), 50.7% (95% CI, 48.5–52.9%), 47.7% (95% CI, 45.4–49.9%), and 90.6% (95% CI, 88.8–92.2%), respectively.
Conclusions Although there are limitations and requirements for applying automated histopathological classification of gastric biopsy specimens in the clinical setting, the results of the present study are promising.

Keywords Gastric biopsy · Automated image analysis · Artificial intelligence · Histopathological classification

Electronic supplementary material The online version of this article (doi:10.1007/s10120-017-0731-8) contains supplementary material, which is available to authorized users.

✉ Hiroshi Yoshida
hiroyosh@ncc.go.jp

[1] Division of Pathology and Clinical Laboratories, National Cancer Center Hospital, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan

[2] Epidemiology and Prevention Group, Center for Public Health Sciences, National Cancer Center, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan

[3] Medical Solutions Division, NEC Corporation, 5-7-1 Shiba, Minato-ku, Tokyo 108-8001, Japan

[4] Space System Division, NEC Corporation, 10, Nisshin-cho 1-Chome, Fuchu, Tokyo 183-8501, Japan

[5] Department of Machine Learning, NEC Laboratories America, 4 Independence Way, Suite 200, Princeton, NJ 08540, USA

[6] Division of Molecular Pathology, National Cancer Center Research Institute, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan

[7] Division of Pathology, Research Center for Innovative Oncology, National Cancer Center, 6-5-1, Kashiwa, Chiba 277-8577, Japan

Introduction

Digital pathology techniques including automated image analysis have been developed and widely utilized in research and in the practice of surgical pathology [1, 2]. For supporting diagnostic procedures, various novel devices have been reported to be effective, including an automated screening system for cytopathology [3], automated analysis for immunohistochemical biomarkers [4], and

Table 1 The revised Vienna classification, Japanese "Group classification," and classification by e-Pathologist

The revised Vienna classification	Description	Japanese "Group classification"	Classification by e-Pathologist
Category 1	Negative for neoplasia/dysplasia	Group 1	Negative
Category 2	Indefinite for neoplasia/dysplasia	Group 2	Caution
Category 3	Low-grade adenoma/dysplasia	Group 3	Caution
Category 4.1	High-grade adenoma/dysplasia	Group 4	Positive
Category 4.2	noninvasive carcinoma	Group 5	Positive
Category 4.3	Suspicion for invasive carcinoma	Group 5	Positive
Category 5.1	Intramucosal carcinoma	Group 5	Positive
Category 5.2	Submucosal invasive carcinoma	Group 5	Positive
Category X	Inadequate specimen for diagnosis	Group X	Unclassifiable

automated morphological analysis and classification for hematoxylin and eosin (H&E)-stained slides [5–7]. However, there has been no report on automated image analysis and histological classification in clinical settings for gastrointestinal cancers.

The need for automated image analysis of gastrointestinal cancers has been increasing. Gastric cancer and colorectal cancer are among the five major cancers in Japan [8]; thus, a large number of endoscopically obtained specimens are being submitted for pathological analysis. This considerable workload for surgical pathologists needs to be reduced; automated screening for negative specimens that do not require the review of a pathologist could be effective. Moreover, application of automated image analysis is expected to contribute to the quality control of routine pathological diagnosis.

NEC Corporation has developed the e-Pathologist image analysis software that can classify digitized histological images of gastric biopsy specimens into three categories that correspond to carcinoma or suspicion of carcinoma (positive), adenoma or suspicion of a neoplastic lesion (caution), and no malignancy (negative). However, the validity of this software analysis in routine pathological practice remains unclear.

The aim of the present study was to evaluate the accuracy of the classification of the e-Pathologist image analysis software and clarify the requirements for using an automated screening system in clinical settings.

Materials and methods

Patient selection, tissue section preparation, and pathological diagnosis

The study was conducted in accordance with the Declaration of Helsinki, and with the approval of the Institutional Review Board of the National Cancer Center, Tokyo, Japan. We consecutively collected a total of 3062 gastric biopsy specimens between January 19 and April 30 in 2015 at the National Cancer Center (Tsukiji and Kashiwa campus). The specimens were fixed in 10% buffered formalin and embedded in paraffin. Each block was sliced into 4-μm-thick sections. Routine hematoxylin and eosin (H&E) staining was performed for each slide using an automated staining system.

At least two experienced gastrointestinal pathologists evaluated each slide for pathological diagnosis according to the Japanese classification [9] and the revised Vienna classification [10] for routine clinical practice. They determined the final consensus diagnosis of each case. The Japanese group classification was applied to only endoscopic biopsy materials and epithelial tissue. In brief, this classification defines group X as an inappropriate material for histological diagnosis, Group 1 as a normal tissue or nonneoplastic lesion tissue, Group 2 as a material in which a diagnosis of a neoplastic or nonneoplastic lesion is difficult, Group 3 as an adenoma tissue, Group 4 as a neoplastic lesion tissue that is suspected to be carcinoma, and Group 5 as a carcinoma tissue. The revised Vienna classification has been widely accepted and consists of five categories. The Vienna classification corresponds to the Japanese group classification as follows: Category 1 is negative for neoplasia/dysplasia (Group 1); Category 2 is indefinite for neoplasia/dysplasia (Group 2); Category 3 represents noninvasive low-grade neoplasia/dysplasia (Group 3); Category 4.1 represents high-grade adenoma/dysplasia (Group 4); and Category 4.2–5.2 represents noninvasive carcinoma to submucosal carcinoma, or beyond (Group 5). The correspondence of each category or group and the final output of e-Pathologist are summarized in Table 1.

Digital image acquisition

After data anonymization and setting image acquisition parameters, such as magnification and autofocusing mode,

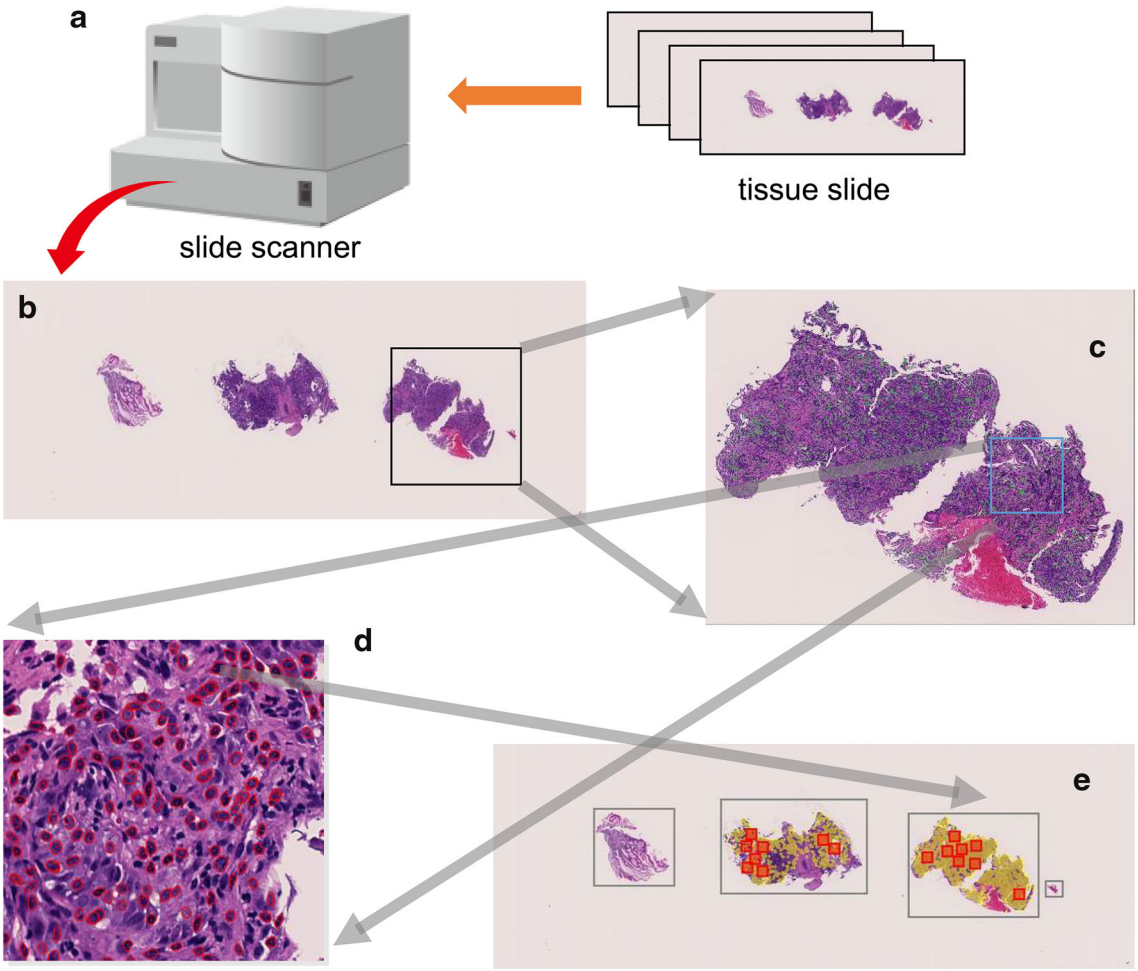

Fig. 1 Analysis flow of the e-Pathologist. **a** Image scanning using a slide scanner. **b** Tissue mapping at ×2.5. **c** Structural analysis at ×10. **d** Nuclear analysis at ×20. **e** Results of the classification analysis are shown in *colored rectangles*

all slides were automatically scanned using the virtual slide scanner NanoZoomer (Hamamatsu Photonics, Shizuoka, Japan) at 40× magnification (0.23 μm/pixel). For the image database, the NDP serve slide image system of Hamamatsu Photonics was used. During the image collection and analysis procedure, the researchers and statistician (T.S.) were blind to all the diagnoses made by the human pathologists.

Image data acquisition

Before the evaluation using test slides, a machine learning algorithm was trained using a large set of H&E-stained gastric tissue sections on standard glass slides.

Cancer detection procedure

The procedure for detection of a cancerous areas in a given whole-slide image is shown in Fig. 1. Briefly, each slide was scanned, resulting in a single whole-slide image that was the input for automated analysis training. The whole-slide images had multiple resolution layers that enabled access to a variety of images acquired at 1.25×, 2.5×, 5×, 10×, and 40× objective lens magnification. The first step of analysis was to identify the tissue regions at 1.25× magnification. The color distribution of the tissue was also analyzed. The tissue area was then divided into several rectangular regions of interest (ROIs). Each ROI was analyzed at a different magnification, i.e., 10× and 20×, depending on the target features to be analyzed (structural or nuclear features). For this analysis the ROI size was 1024 × 1024 pixels, and we obtained an average of 20 ROIs per tissue. After the feature extraction, all ROIs were classified as positive (cancer) or negative (benign) using a trained classifier. The ROI classifier assigns a real number t, in the range $[-1.0; 1.0]$, where a value of 1.0 indicates a positive (cancer) ROI and a value of -1.0 indicates a negative (benign) ROI. The t value can also be interpreted as a confidence level where values close to 0 indicate a low level of confidence.

A positive or negative tissue-level classification was based on the following rule: if $N(t_1)/N_0 > t_2$, the tissue is classified to positive, otherwise negative, where N_0 is the number of ROIs, and $N(t_1)$ is the number of ROIs with $t > t_1$. Real values t_1 and t_2 were optimized using a validation set that was not used for the training of the ROI classifier.

Quantitative characterization of histopathological features

To conduct a positive or negative classification using a machine learning algorithm, it was necessary to extract several types of histopathological features. These features were divided into two categories: high-magnification and low-magnification features.

High-magnification features characterized the nuclear morphology and texture of an ROI. The contours of nuclei were traced by following the elliptical boundaries of eosin-stained and hematoxylin-stained pixels using a dynamic programming approach. After the extraction of nuclear contours, several features could be readily computed. To characterize nuclear morphology, we considered the mass (number of pixels bounded by the contour) and long-axis length (after an elliptical fit of the contour points); for texture, we considered the standard deviation (variance) of the hematoxylin color channel within the area bounded by the contours. Because an ROI contains several hundred nuclei contours, we needed to extract statistical measures for the entire ROI. Experiments have shown that it is beneficial to independently extract statistics for small and large nuclei (based on their pixel mass). Thus, we calculated the mean, standard deviation, and 85th percentile of three features (two morphological, and one texture) for two nuclei groups (small and large), resulting in $2 \times 3 \times 3 = 18$ features. We also included the total number of small and large nuclei within the ROI, resulting in 20 high-magnification features per ROI.

Low-magnification features characterized the global H&E stain distribution within an ROI and the appearance of blood cells and gland formation. We considered the proportion of tissue pixels that belong to the categories of hematoxylin (H), eosin (E), and blood (B). To obtain this classification, we first trained a support vector regression (SVR) model to predict the color vector (red, green, or blue pixel intensity) of H, E, or B tissue pixels given the overall color histogram of the ROI. A set of training ROIs were labeled for H, E, B colors using an interactive color picker, and the SVR classifier was trained in a standard supervised fashion. The ROI pixels were then classified based on their proximity to these color vectors, resulting in four features (%H, %E, %B, and H/E). Gland formations are difficult to characterize and extract for analysis. Therefore, we use a data-driven machine learning approach to train a model to extract gland and duct formations. Convolutional neural networks (CNN) [11] are well suited for this task and can be trained directly from the RGB pixels on a set of images where glands have been traced to provide the training label. We extracted two features within the ROI, the number of glands and the proportion of pixels belonging to a gland. The number of glands was obtained by counting the number of blobs returned by connected component analysis on the binarized CNN output. The total number of low- and high-magnification features is thus $20 + 4 + 2 = 26$.

The procedure for detection described here does not detect nonepithelial malignancies such as lymphomas and carcinoid tumors. The possibility of such nonepithelial cancers was determined using a rule-based classifier based on tissue-wide features (i.e., nuclear density and the ratio of large, medium, and small nuclei). The rule-based classifier was also trained to detect carcinomas that are difficult to detect based on the 26 features just described. The specimens that the system analyzed as suspicious for carcinoma using the rule-based classifier were categorized using the term "caution." Thus, the system classifies a given tissue image into three categories: positive, negative, or caution. When the quality of the input image is inappropriate for analysis as a result of bad staining (i.e., too weak or too strong) or bad imaging (i.e., blurring), the system excludes these images as "unclassifiable." This decision is made automatically by a rule-based classifier based on the overall color histogram and a blurriness detector based on FFT (Fast Fourier Transform).

Training of the tissue classifier

The difficulty in training a tissue-level classifier is that a whole-tissue image typically only exhibits cancerous features on a small part of the tissue (usually a few ROIs). It is rare that the entire tissue is visibly cancerous. For a standard supervised machine learning approach to work, a panel of experts would need to have each ROI in the training set labeled as positive or negative. This effort is feasible for small datasets of only a few hundred tissues; however, in our study, we analyze tens of thousands of tissues that expert pathologists have evaluated as ground-truth specimens at the whole-tissue level (not at the ROI level).

To overcome this difficulty, we used multi-instance learning (MIL) [12] for training a multilayer neural network (MLNN) model. In MIL, a whole tissue is represented by a "bag" of instances (ROIs) and a single label (the whole-tissue ground-truth label). When training a positive-labeled tissue example, only the instance (ROI)

generating the largest forward response from the MLNN is back propagated to adjust the model, accounting for the fact that only a few ROI within the tissue are positive. In this way, we were able to train a robust model using a large training dataset of 26,595 tissues with only tissue-level annotations [13].

Statistical analysis

We compared the classification results of human pathologists and those of the e-Pathologist software. To compare the results of the three-tier classification, agreement was assessed as the percent agreement and kappa coefficients (κ) [14]. Kappa coefficients ranged from 0.00 to 1.00 and were interpreted descriptively as follows: poor $\kappa < 0.20$, fair $\kappa = 0.20$–0.40, moderate $\kappa = 0.40$–0.60, good $\kappa = 0.60$–0.80, and very good $\kappa = 0.80$–1.00.

To calculate the sensitivity, specificity, positive predictive value, and negative predictive value of the images judged as "negative" by e-Pathologist, we dichotomized the three-tier classification results as negative for negative specimen images and as non-negative for positive, caution, and unclassifiable specimen images. Each parameter was defined as follows: (1) sensitivity: true non-negative/(true non-negative + false negative); (2) specificity: true negative/(true negative + false non-negative); (3) positive predictive value: true non-negative/(true non-negative + false non-negative); and (4) negative predictive value: true negative/(true negative + false negative). All statistical analyses were performed using JMP 10.0.0 software (SAS Institute, Cary, NC, USA).

Results

Final diagnoses by human pathologists

The details of the final diagnoses by human pathologists are summarized in Table 2. Of the 3062 cases, 66.6% were diagnosed as Group 1 cases (negative for a tumor) and the remaining cases showed an abnormal finding.

Comparison of three-tier classification results

We compared the three-tier classifications of the human pathologists and e-Pathologist. The results are shown in Table 3. The overall concordance rate was 55.6% (1702/3062). The kappa coefficient was 0.28 (95% CI, 0.26–0.30, fair agreement). For the negative biopsy specimens, the concordance rate was 90.6% (1033/1140), but was less than 50% for the positive biopsy specimens. Of the 3062 specimens, e-Pathologist regarded 215 (7%) specimens as unclassifiable.

Table 2 Summary of the final diagnoses by human pathologists

Final diagnosis	Number (%)
Group 1	2039 (66.6)
Group 2	26 (0.9)
Group 3	173 (5.7)
Group 4	22 (0.7)
Group 5	729 (23.8)
Others	73 (2.4)
Lymphoma	39 (1.3)
Carcinoid	9 (0.3)
GIST	9 (0.3)
Atypical lymphoid lesion	8 (0.3)
Plasmacytoma	3 (0.1)
Leiomyoma	2 (0.1)
Anisakiasis	1 (0.03)
Insufficient material	2 (0.1)
Total (%)	3062 (100)

Table 3 Comparison of three-tier classification results between human pathologists and e-Pathologist

	Human pathologists			Concordance rate
	Positive	Caution	Negative	
e-Pathologist				
Positive	658	118	726	43.8% (658/1502)
Caution	22	11	172	5.4% (11/205)
Negative	67	40	1033	90.6% (1033/1140)
Unclassifiable	67	40	108	0% (0/215)
Total	814	209	2039	55.6% (1702/3062)

Comparison of two-tier classification results

Our primary interest was to determine whether e-Pathologist could accurately screen specimens without the need for further human pathologist review. Therefore, we compared the two-tier classifications of human pathologists and e-Pathologist. The results are summarized in Table 4. The sensitivity, specificity, positive predictive value, and negative predictive value were 89.5% (95% CI, 87.5–91.4%), 50.7% (95% CI, 48.5–52.9%), 47.7% (95% CI, 45.4–49.9%), and 90.6% (95% CI, 88.8–92.2%), respectively.

Analysis of false-negative results

False-negative classifications of the e-Pathologist software are a serious error because a human pathologist might not review such specimens in the clinical setting. Therefore, we identified the causes of false-negative classifications in a detailed case review, as summarized in Table 5 and

Table 4 Comparison of two-tier classification results between human pathologists and e-Pathologist

	Human pathologist		
	Negative	Non-negative	Total
e-Pathologist			
Negative	1033	107	1140
Non-negative	1006	916	1922
Total	2039	1023	3062

Supplementary Table 1. Of the 1140 e-Pathologist specimens classified as negative, 107 (9.4%) had an abnormal finding. Representative false-negative results are shown in Fig. 2 (examples of false-positive cases are shown in Supplementary Fig. 1). Well-differentiated noninvasive neoplasia (low- and high-grade adenoma/dysplasia and carcinoma in situ) and poorly differentiated adenocarcinoma accounted for approximately 75% of the false-negative cases. Furthermore, 11% of the false-negative results were for lymphoid lesion specimen images. Of all 47 lymphoma and atypical lymphoid lesion specimen images, 12 (26%) were classified as "Negative."

After the final review of all the discordant cases, we confirmed e-Pathologist did not identify any neoplasms missed by human pathologists.

Discussion

The present study is the first attempt to investigate the efficacy of automated image analysis software (e-Pathologist) for screening gastric biopsy specimens. A total of 3062 specimens were digitized and analyzed, of which one-third showed abnormal findings. We compared the classification results between human pathologists and e-Pathologists. Although the overall concordance rate was as low as 55.6% for the three-tier classification, e-Pathologist could accurately identify 90.6% of negative specimens. The results of this first large-scale study of automated image analysis for gastric biopsy specimen are encouraging. However, several requirements and limitations should be considered for the use of e-Pathologist in daily clinical settings.

In the present study, the prevalence of abnormal findings exceeded 33%, and included cases with mesenchymal tumor or a lymphoid malignancy other than epithelial neoplasm. Thus, these consecutively collected specimens exhibited a variety of diseases, reflecting the actual clinical setting.

The overall concordance rate for the three-tier classification was 55.6%, and the kappa coefficient was 0.28, indicating fair agreement. The level of diagnostic agreement among pathologists is substantially higher. Nakhleh et al. summarized 84 relevant studies and reported a major discrepancy median of 6.3% (1.9–10.6%, 25th–75th percentile) [15]. Studies on diagnostic discrepancies in gastrointestinal pathology reported discrepancy rates as low as 1.2–3.1% [16–18]. If we intended to use e-Pathologist in a clinical setting, without supervision by a human pathologist, discrepancy levels at least as low as those reported previously should be required. In contrast, approximately 90% of negative cases were accurately predicted. This result is promising for the use of e-Pathologist as a screener of specimens without further human pathologist review. In the area of cytopathology, the AutoPap Primary Screening

Table 5 Detailed final diagnosis and classification by e-Pathologist

Final diagnosis	Classification by e-Pathologist				Total
	Positive	Caution	Negative	Unclassifiable	
Group 1	726	172	1033	108	2039
Group 2	18	2	4	2	26
Group 3	96	7	32	38	173
Group 4	18	1	3	0	22
Group 5	595	17	55	62	729
Lymphoma	25	3	8	3	39
GIST	9	0	0	0	9
Carcinoid	7	1	1	0	9
Atypical lymphoid	3	1	4	0	8
Plasmacytoma	1	0	0	2	3
Leiomyoma	2	0	0	0	2
Anisakisasis	1	0	0	0	1
Insufficient	1	1	0	0	2
Total	1502	205	1140	215	3063

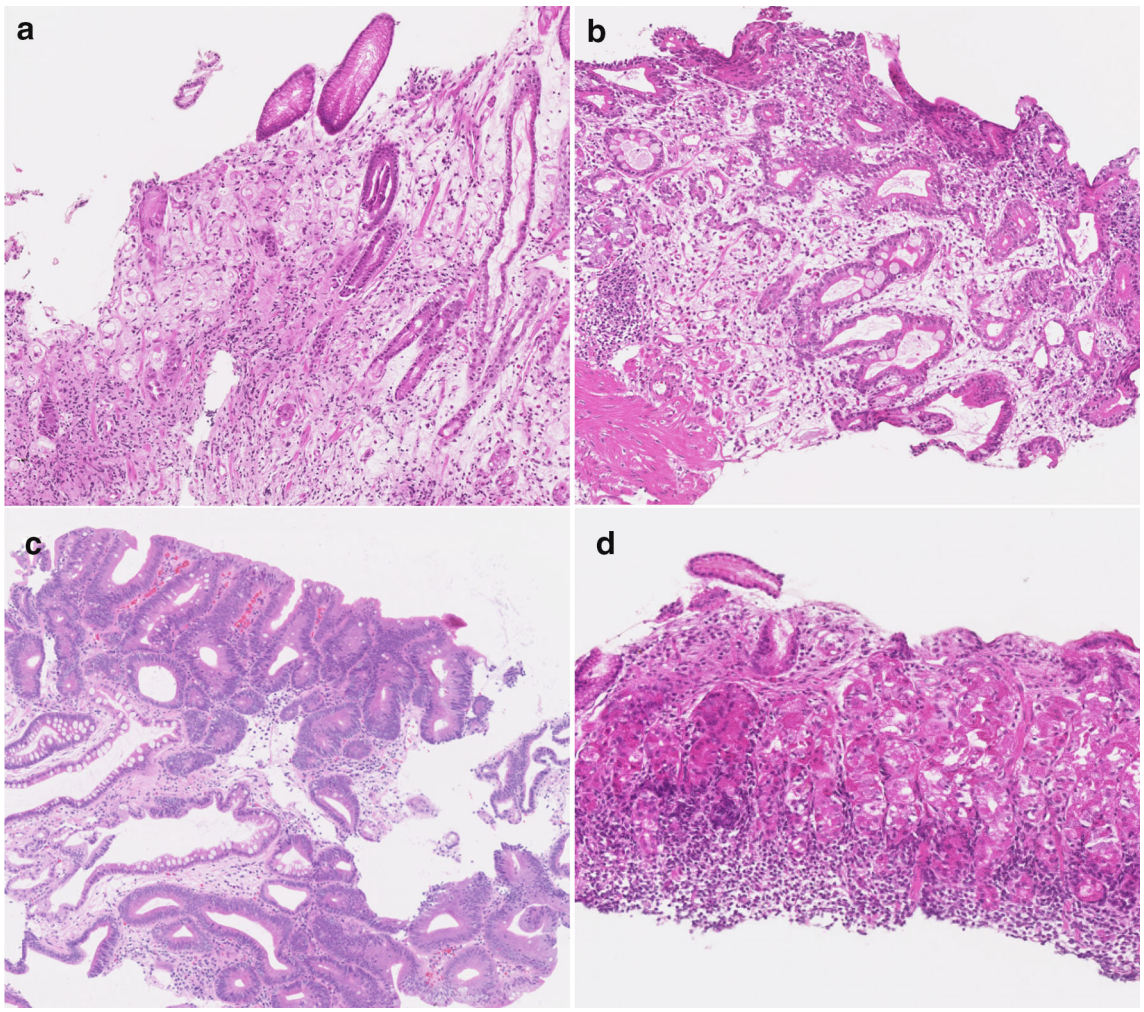

Fig. 2 Representative images of hematoxylin and eosin (H&E)-stained tissue from false-negative cases. **a** Poorly differentiated adenocarcinoma and signet-ring cell adenocarcinoma. **b** Well-to-moderate differentiation of tubular adenocarcinoma. **c** Tubular adenoma, intestinal type. **d** MALT lymphoma. The e-Pathologist software classified all the cases herein as negative

System has received the approval of the United States Food and Drug Administration for the initial screening and quality control of cervical cytology slides. This system shows statistically superior abnormality detection sensitivity to that of the current standard practice of manual screening. For gastric biopsy specimens, there have been no comparable data on the sensitivity and specificity of abnormal finding screening. However, studies of the diagnostic error in surgical pathology may serve as useful references. Based on the reported gastrointestinal pathology diagnostic disagreement rate of 1.2–3.1% [16–18], a false-negative rate of 1–5% might be considered as an interim goal for primary screening. However, there is no consensus on the acceptable false-negative rate for regular pathological diagnosis. Various factors complicate this problem, such as the severity or curability of the disease, the maturity of the healthcare system, including pathologist accessibility, and the cultural background of each country. Nevertheless, with a false-negative rate similar to those reported for human pathologists, it may be acceptable at present to use automated image analysis for routine practice.

In the present study, the prevalence of abnormal findings exceeded 33%, which was mainly the result of the character of specialized cancer center hospitals from which the data were obtained. This relatively high prevalence of abnormal findings increased the absolute number of false-negative cases. For a normal prevalence value of 1% for screening the general population using upper gastrointestinal endoscopy [19], the e-Pathologist test, with its 90% sensitivity and 50% specificity for detection, would perform at 99.8% negative predictive value and 1.7% positive predictive value.

To improve the screening ability of e-Pathologist, the false-negative classification rate should be reduced. In the present study, 86% of false-negative cases involved low-

grade noninvasive epithelial neoplasia, small amounts of poorly differentiated adenocarcinoma/signet-ring cell carcinoma scattering in lamina propria mucosae, and lymphoid lesions, such as MALT lymphoma. These lesions are difficult to diagnose and exhibit lower interobserver reproducibility for pathological diagnosis [20, 21]. Machine learning models trained specifically for these lesions may help reduce the e-Pathologist false-negative classification rate.

A significant number of cases were deemed "unclassifiable" by e-Pathologist, totaling 7% of all specimens. The reasons for such classification included poor staining (too weak or too strong) or imaging problems (blurriness). In clinical settings, the quality of H&E staining or scanned digital images can vary, and human pathologists correctly diagnosed those same specimens. Nevertheless, rejection of difficult-to-classify specimens is a reasonable way to increase robustness until a sufficient number of such difficult cases can be collected to train specific models.

In routine practice, human pathologists need to consider not only neoplastic disease but also nonneoplastic entities, including infectious or inflammatory disease. For these specimens, a pathological report of negative for neoplasia/dysplasia is insufficient and inappropriate. For example, characterization of gastritis, intestinal metaplasia. and atrophy are key reportable features, especially in countries with a low prevalence of *Helicobacter pylori* infection. Pathologists must perform further review of these slides and make relevant comments on clinical diagnosis. At present, e-Pathologist cannot meet these requirements, and further improvements should therefore involve the ability to recognize nonneoplastic disease. In addition, the present form of analysis was based on only visible morphological, textual, and color features. The process of making a pathological diagnosis requires the integration of a variety of clinical information and background knowledge. In cases involving a major discrepancy between clinical diagnosis and histomorphological findings, most pathologists would carefully consider the cause of the discrepancy and perform appropriate actions, such as communicating with an endoscopist, ordering deeper sections, or checking for misidentification of the specimen. Therefore, for the benefit of the patient, we must recognize this essential difference between human pathologist diagnosis and AI-based automated image classification and carefully consider the regulations and requirements for the use of automated image analysis in the clinical setting.

In conclusion, although there are some limitations and requirements for using automated histopathological classification of gastric biopsy specimens in clinical settings, the present study shows promising results. Further improvements in machine learning to reduce false-negative classification may help realize the potential of automated screening to aid pathologists in the not so distant future.

Acknowledgements The present investigation received the support of the National Cancer Center Research and Development Fund (#26-A-7). We thank Sachiko Fukuda, Kumiko Yamabe, and Akihisa Kondo for their technical assistance. We also thank Editage (http://www.editage.jp) for English language editing.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical standards All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1964 and later versions. Informed consent or substitute for it was obtained from all patients for being included in the study.

References

1. Al-Janabi S, Huisman A, Van Diest PJ. Digital pathology: current status and future perspectives. Histopathology (Oxf). 2012;61(1):1–9.
2. Park S, Parwani AV, Aller RD, Banach L, Becich MJ, Borkenfeld S, et al. The history of pathology informatics: a global perspective. J Pathol Inform. 2013;4:7.
3. Wilbur DC, Prey MU, Miller WM, Pawlick GF, Colgan TJ. The AutoPap system for primary screening in cervical cytology. Comparing the results of a prospective, intended-use study with routine manual practice. Acta Cytol. 1998;42(1):214–20.
4. Stalhammar G, Fuentes Martinez N, Lippert M, Tobin NP, Molholm I, Kis L, et al. Digital image analysis outperforms manual biomarker assessment in breast cancer. Mod Pathol. 2016;29(4):318–29. doi:10.1038/modpathol.2016.34 **(Epub Feb 26)**.
5. Frydenlund A, Eramian M, Daley T. Automated classification of four types of developmental odontogenic cysts. Comput Med Imaging Graph. 2014;38(3):151–62.
6. El Hallani S, Guillaud M, Korbelik J, Marginean EC. Evaluation of quantitative digital pathology in the assessment of Barrett esophagus-associated dysplasia. Am J Clin Pathol. 2015;144(1):151–64. doi:10.1309/AJCPK0Y1MMFSJDKU.
7. Vanderbeck S, Bockhorst J, Komorowski R, Kleiner DE, Gawrieh S. Automatic classification of white regions in liver biopsies by supervised machine learning. Hum Pathol. 2014;45(4):785–92. doi:10.1016/j.humpath.2013.11.011 **Epub Nov 26**.
8. Katanoda K, Hori M, Matsuda T, Shibata A, Nishino Y, Hattori M, et al. An updated report on the trends in cancer incidence and mortality in Japan, 1958–2013. Jpn J Clin Oncol. 2015;45(4):390–401.
9. Sano T, Aiko T. New Japanese classifications and treatment guidelines for gastric cancer: revision concepts and major revised points. Gastric Cancer. 2011;14(2):97–100.
10. Schlemper RJ, Riddell RH, Kato Y, Borchard F, Cooper HS, Dawsey SM, et al. The Vienna classification of gastrointestinal epithelial neoplasia. Gut. 2000;47(2):251–5.
11. LeCun Y, Bottou L, Orr GB, Müller K-R. Efficient BackProp. In: Orr GB, Müller K-R, editors. Neural networks: tricks of the trade. Berlin: Springer; 1998. p. 9–50.

12. Dietterich TG, Lathrop RH. Solving the multiple instance problem with axis-parallel rectangles. Artif Intell. 1997;89(1–2):31–71.
13. Cosatto E, Laquerre P-F, Malon C, Graf H-P, Saito A, Kiyuna T, et al. Automated gastric cancer diagnosis on H&E-stained sections; training a classifier on a large scale with multiple instance machine learning. Proc SPIE–Int Soc. Opt Eng. 2013;8676:05. doi:10.1117/12.2007047.
14. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther. 2005;85(3):257–68.
15. Nakhleh RE, Nose V, Colasacco C, Fatheree LA, Lillemoe TJ, McCrory DC, 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(1):29–40.
16. Kronz JD, Westra WH, Epstein JI. Mandatory second opinion surgical pathology at a large referral hospital. Cancer (Phila). 1999;86(11):2426–35.
17. Renshaw AA, Gould EW. Measuring errors in surgical pathology in real-life practice: defining what does and does not matter. Am J Clin Pathol. 2007;127(1):144–52.
18. Renshaw AA, Gould EW. Comparison of disagreement and amendment rates by tissue type and diagnosis: identifying cases for directed blinded review. Am J Clin Pathol. 2006;126(5):736–9.
19. Sugano K. Screening of gastric cancer in Asia. Best Pract Res Clin Gastroenterol. 2015;29(6):895–905.
20. Ahn S, Park do Y. Practical points in gastric pathology. Arch Pathol Lab Med. 2016;140(5):397–405.
21. El-Zimaity HM, Wotherspoon A, de Jong D, Houston MALT lymphoma Workshop. Interobserver variation in the histopathological assessment of MALT/MALT lymphoma: towards a consensus. Blood Cells Mol Dis. 2005;34(1):6–16.