Feedback from artificial intelligence improved the learning of junior endoscopists on histology prediction of gastric lesions

OPEN ACCESS

Authors

Thomas K.L. Lui[1], Kenneth K.Y. Wong[2], Loey L.Y. Mak[1], Elvis W.P. To[1], Vivien W.M. Tsui[1], Zijie Deng[3], Jiaqi Guo[3], Li Ni[3], Michael K.S. Cheung[1,3], Wai K. Leung[1]

Institutions

1 Department of Medicine, Queen Mary Hospital, University of Hong Kong, Hong Kong, China
2 Department of Computer Science, University of Hong Kong, Hong Kong, China
3 Department of Medicine, University of Hong Kong-Shenzhen Hospital, Shenzhen, China

Corresponding author

Wai K. Leung, Department of Medicine, Queen Mary Hospital, University of Hong Kong, Hong Kong, China
Fax: +852 2816 2863
waikleung@hku.hk

ABSTRACT

Background and study aims Artificial intelligence (AI)-assisted image classification has been shown to have high accuracy on endoscopic diagnosis. We evaluated the poten-tial effects of use of an AI-assisted image classifier on train-ing of junior endoscopists for histological prediction of gas-tric lesions.

Methods An AI image classifier was built on a convolution-al neural network with five convolutional layers and three fully connected layers A Resnet backbone was trained by 2,000 non-magnified endoscopic gastric images. The inde-pendent validation set consisted of another 1,000 endo-scopic images from 100 gastric lesions. The first part of the validation set was reviewed by six junior endoscopists and the prediction of AI was then disclosed to three of them (Group A) while the remaining three (Group B) were not provided this information. All endoscopists reviewed the second part of the validation set independently.

Results The overall accuracy of AI was 91.0% (95% CI: 89.2–92.7%) with 97.1% sensitivity (95% CI: 95.6–98.7%), 85.9% specificity (95% CI: 83.0–88.4%) and 0.91 area un-der the ROC (AUROC) (95% CI: 0.89–0.93). AI was superior to all junior endoscopists in accuracy and AUROC in both validation sets. The performance of Group A endoscopists but not Group B endoscopists improved on the second vali-dation set (accuracy 69.3% to 74.7%; $P = 0.003$).

Conclusion The trained AI image classifier can accurately predict presence of neoplastic component of gastric le-sions. Feedback from the AI image classifier can also hasten the learning curve of junior endoscopists in predicting his-tology of gastric lesions.

Introduction

Gastric cancer is the fifth most common cancer and accounts for more than 800,000 deaths worldwide each year [1]. Early detection and accurate characterization of gastric neoplastic lesions during endoscopy is of paramount importance because the prognosis of early gastric cancer is excellent [2,3]. How-ever, early gastric neoplastic lesions are usually subtle and easi-ly missed [4]. Use of optical magnified endoscopy in combina-tion with chromoendoscopy or image-enhanced endoscopy such as narrow-band imaging (NBI) has been suggested to help differentiate and characterize early gastric lesions by en-hancing the microsurface and microvascular pattern. In partic-ular, irregular microsurface and microvascular pattern under NBI examination was associated with presence of intraepithelial neoplasia [5–9]. Nevertheless, this kind of endoscopic diagnos-tic skill requires a considerable amount of training and experi-ence, which may not be readily available in most endoscopy units.

Absent reliable histological prediction of endoscopic gastric lesions, the gold standard for diagnosis of gastric lesions usually requires multiple biopsies or even total en bloc resection, as a

Lui Thomas K.L. et al. Feedback from artificial… Endoscopy International Open 2020; 08: E139–E146

E139

single biopsy may miss the most advanced pathology of a lesion. However, processing of multiple biopsies is costly and complete excision of large gastric lesions is technical challenging [10]. Sampling error also can produce false-negative results [11]. With rapid development of artificial intelligence (AI) in endoscopy, a pilot study has shown the possibility of using AI for accurate detection of early gastric lesions [12]. A recent article also showed the potential of AI in predicting depth of invasion of gastric lesions [13].

So far, however, there are no data on investigations specifically of the role of AI in training of junior endoscopists. In this study, we assessed the role of AI in training junior endoscopists in predicting histology of endoscopic gastric lesions.
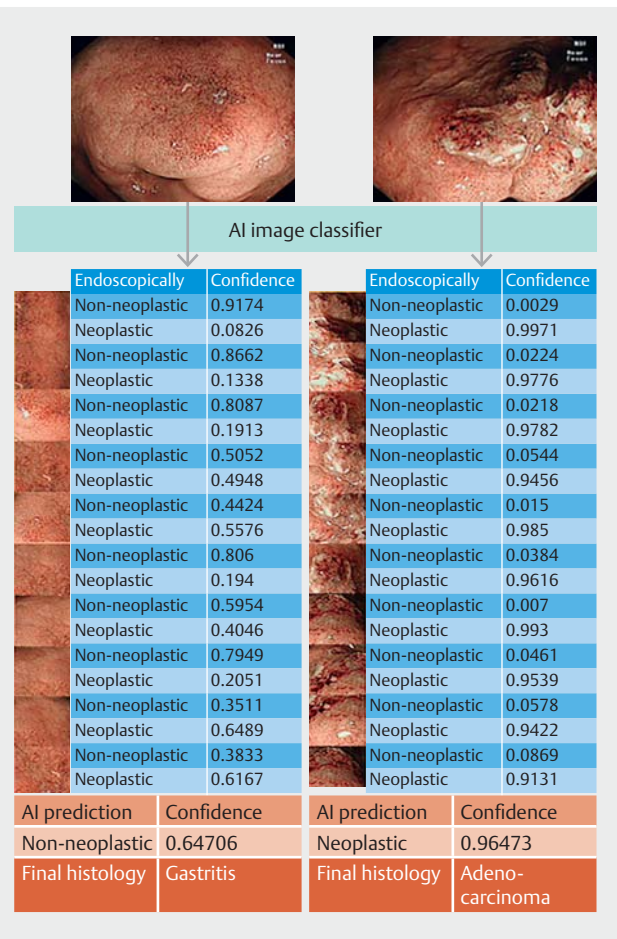
Method

Setting

The study was conducted in the Integrated Endoscopy Center of the Queen Mary Hospital of Hong Kong, which is a major regional hospital serving the Hong Kong West Cluster and a university teaching hospital. The study protocol was approved by the Institutional Review Board of the Hospital Authority Hong Kong West Cluster and the University of Hong Kong.

All baseline endoscopies were performed with non-optical magnifying gastroscope (GIF-HQ290 model and CV-290 video system, Olympus, Tokyo, Japan).

In this study, we included only gastric lesions with Paris Classification type 0-IIa, IIb, IIc or Is. In addition to elevated lesions, subtle mucosal changes or ulcer scars that have similar shapes to IIc lesions were also included. Still endoscopic images were retrieved from the electronic patient record system or the archive endoscopic video system of our endoscopy unit. Image resolution was at least 720×526 pixels and images were obtained under NBI. NBI was used as our previous study had demonstrated its superiority over white light for AI interpretation [14]. The gold standard was the final gastric pathology which was based on multiple biopsies or total endoscopic resection of the lesion, and classified according to the WHO classification [15]. Neoplastic lesions were defined pathologically as presence of intraepithelial neoplasia (dysplasia) or adenocarcinoma in the most advanced histology of a lesion. Non-neoplastic lesions were defined as absence of intraepithelial neoplasia (dysplasia) or adenocarcinoma in any part of a lesion.

Building the AI image classifier and training set

An AI image classifier was built on a convolutional neural network (CNN) with five convolutional layers and three fully connected layers by using endoscopic images of gastric lesions obtained between January 2013 and December 2016. The AI image classifier was based on a pre-trained ResNet CNN backbone. All the training images were pre-screened by an experienced endoscopist (TKLL), who had performed more than 4,000 image-enhanced upper endoscopies with NBI. Multiple images per lesion were obtained in the training set by image augmentation including rotation, flipping, and reversing to expand the training set. The region of interest (ROI) within the endoscopic images (300×300 pixels) was randomly highligh-
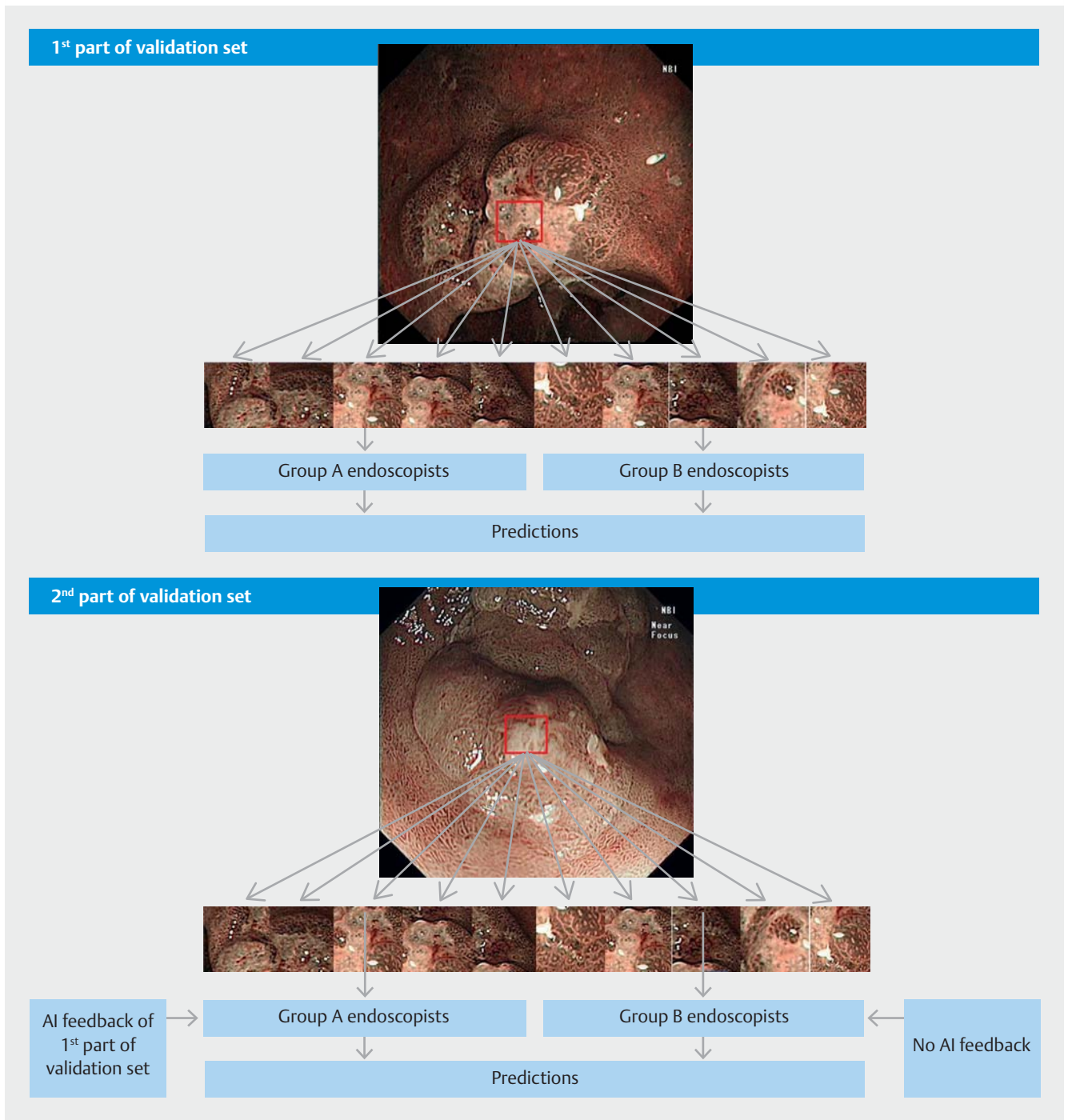

AI image classifier			
	Endoscopically	Confidence		Endoscopically	Confidence

	Endoscopically	Confidence
	Non-neoplastic	0.9174
	Neoplastic	0.0826
	Non-neoplastic	0.8662
	Neoplastic	0.1338
	Non-neoplastic	0.8087
	Neoplastic	0.1913
	Non-neoplastic	0.5052
	Neoplastic	0.4948
	Non-neoplastic	0.4424
	Neoplastic	0.5576
	Non-neoplastic	0.806
	Neoplastic	0.194
	Non-neoplastic	0.5954
	Neoplastic	0.4046
	Non-neoplastic	0.7949
	Neoplastic	0.2051
	Non-neoplastic	0.3511
	Neoplastic	0.6489
	Non-neoplastic	0.3833
	Neoplastic	0.6167

	Endoscopically	Confidence
	Non-neoplastic	0.0029
	Neoplastic	0.9971
	Non-neoplastic	0.0224
	Neoplastic	0.9776
	Non-neoplastic	0.0218
	Neoplastic	0.9782
	Non-neoplastic	0.0544
	Neoplastic	0.9456
	Non-neoplastic	0.015
	Neoplastic	0.985
	Non-neoplastic	0.0384
	Neoplastic	0.9616
	Non-neoplastic	0.007
	Neoplastic	0.993
	Non-neoplastic	0.0461
	Neoplastic	0.9539
	Non-neoplastic	0.0578
	Neoplastic	0.9422
	Non-neoplastic	0.0869
	Neoplastic	0.9131

AI prediction	Confidence
Non-neoplastic	0.64706
Final histology	Gastritis

AI prediction	Confidence
Neoplastic	0.96473
Final histology	Adeno-carcinoma

▶ **Fig. 1** Representative figures of AI image classifier for prediction of histology of sessile gastric lesions.

ted. All images that contained motion artefact, were out of focus, had inappropriate brightness or were covered with mucus were excluded. The final training set consisted of 2,000 ROI images (1,000 ROI images from 170 neoplastic lesions and 1,000 ROI images from 230 non-neoplastic lesions). A total of 10% of the training images were randomly chosen as an internal validation set with 99.5% internal accuracy.

Validation set

The independent validation set consisted of another 1,000 ROI selected from endoscopic images of 100 gastric lesions obtained between January 2017 and January 2019. The ROI within the endoscopic images was selected as described for the training set. To minimize selection bias, 10 ROIs were randomly selected from a single endoscopic image of a lesion. The ROI images were then analyzed by the trained AI image classifier to predict presence of neoplastic lesion (▶ **Fig. 1**).

The validation set was randomly divided into two parts with 500 ROIs in each part. Six junior endoscopists (Endoscopist I to VI) who had performed more than 1,000 upper endoscopies and had undergone special NBI training tutorials on characterizing gastric lesions were asked to comment on whether the ROIs from the first part of the validation set were neoplastic le-

Group A endoscopists

Group B endoscopists

Predictions

AI feedback of 1st part of validation set

Group A endoscopists

Group B endoscopists

No AI feedback

Predictions

▶ **Fig. 2** Study flow.

sions. After the first half of the validation set was reviewed, the prediction result of AI was disclosed to three of them (Group A endoscopists: I, II, III) while the remaining three (Group B endoscopists: IV, V VI) were not provided this information. All six endoscopists then reviewed the second part of the validation set (▶ **Fig. 2**). As a further control, a senior endoscopist who had performed more than 4,000 upper endoscopies with special NBI training on characterizing gastric lesions was also involved in reviewing the validation set.

Statistical analysis

We assumed that AI was superior to an endoscopist and that the accuracy of AI image classifier was 90 %. Assuming a difference of 20 % in accuracy and with a statistical power of 80 % and a two-sided significance level of 0.05, 50 ROI were needed in each study arm. Categorical data were compared by the χ^2-test or Fisher Exact test where appropriate. Numerical data were analyzed by the Student's t-test. Statistical significance was taken as a two-sided $P < 0.05$. For multiple comparisons,

Lui Thomas K.L. et al. Feedback from artificial… Endoscopy International Open 2020; 08: E139–E146

E141

▶ **Table 1** Clinicopathological characteristics of the validation set.

	All	1ˢᵗ part	2ⁿᵈ part	P
Number of lesions	100	50	50	1.00
Mean size (mm)	14.9 mm	15.4 mm	14.7 mm	0.73
Morphology				
IIa or IIa-like	55.0 % (n = 55)	48.0 % (n = 24)	62.0 % (n = 31)	0.23
IIb or IIb-like	22.0 % (n = 22)	18.0 % (n = 9)	26.0 % (n = 13)	0.47
IIc or IIc-like	11.0 % (n = 11)	14.0 % (n = 7)	8.0 % (n = 7)	0.52
Is or Is-like	12.0 % (n = 12)	20.0 % (n = 10)	4.0 % (n = 2)	0.12
Location				
Antrum	71.0 % (n = 71)	72.0 % (n = 36)	70.0 % (n = 35)	1.00
Body	29.0 % (n = 29)	28.0 % (n = 14)	30.0 % (n = 15)	1.00
Histology				
Gastritis	36.0 % (n = 36)	34.0 % (n = 17)	38.0 % (n = 19)	0.84
Intestinal metaplasia	14.0 % (n = 14)	12.0 % (n = 6)	16.0 % (n = 8)	0.77
Hyperplastic	2.0 % (n = 2)	4.0 % (n = 2)	0 % (n = 0)	0.47
Low-grade dysplasia	30.0 % (n = 30)	32.0 % (n = 16)	28.0 % (n = 14)	0.83
High-grade dysplasia	5.0 % (n = 5)	4.0 % (n = 2)	6.0 % (n = 3)	1.00
Adenocarcinoma	13.0 % (n = 13)	14.0 % (n = 7)	12.0 % (n = 6)	1.00
Tumor depth				
Intramucosal	4 % (n = 4)	4 % (n = 2)	4 % (n = 2)	1.00
Submucosal	9 % (n = 9)	10 % (n = 5)	8 % (n = 4)	1.00
Histology subtype				
Well differentiated	6 % (n = 6)	6 % (n = 3)	6 % (n = 3)	1.00
Moderately differentiated	6 % (n = 6)	6 % (n = 3)	6 % (n = 3)	1.00
Poorly differentiated	1 % (n = 1)	2 % (n = 1)	0 % (n = 0)	1.00
AI prediction				
AUROC (95 % CI)	0.92 (0.89–0.93)	0.92 (0.90–0.94)	0.91 (0.89–0.93)	0.53
Accuracy (95 % CI)	91.0 % (89.1–92.7 %)	91.6 % (88.8–93.9 %)	90.4 % (87.5–92.8 %)	0.52

AUROC, area under the receiver operating characteristics curve; CI, confidence interval

the *P* value was adjusted by Bonferroni correction. A two-by-two table was constructed using the predicted and actual outcome to calculate different domains in the diagnostic test with sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and accuracy. Confidence intervals (CIs) used for sensitivity, specificity and accuracy were Clopper-Pearson CIs. CIs for predictive values were the standard logit CIs. All statistical analysis was performed by SPSS statistics software (version 19.0, SPSS, Chicago, Illinois, United States).

Results

Clinicopathological characteristics of the gastric lesions in the validation set are summarized in ▶ **Table 1**. Mean lesion size was 14.9 mm (range: 5 to 40 mm) and 71 were located at the antrum. The majority of the lesions were Paris Type 0 IIa (55.0 %, n = 55) followed by IIb lesion (22.0 %, n = 22), Is lesion (12.0 %, n = 12) and IIc lesion (11.0 %, n = 11). Forty-eight were neoplastic lesions including 13 adenocarcinomas, five high-grade dysplasias and 30 low-grade dysplasias.

▶Table 2 Analysis of the performance of AI according to lesion characteristics.

	Accuracy (95 %CI)	AUROC (95 %CI)
Size		
>10 mm	90.7 % (88.5–92.7 %)	0.90 (0.88–0.92)
≤10 mm	91.9 % (87.4 %-95.2 %)	0.93 (0.89–0.97)
Morphology		
IIa or IIa-like	91.4 % (88.8–93.6 %)	0.92 (0.89–0.94)
IIb or IIb-like	83.6 % (78.0–88.2 %)	0.91 (0.89–0.94)
IIc or IIc-like	98.2 % (96.5–99.9 %)	0.99 (0.97–0.99)
Is or Is-like	95.8 % (90.5–98.6 %)	0.95 (0.91–0.99)
Location		
Antrum	89.2 % (86.8–91.4 %)	0.90 (0.88–0.91)
Body	95.2 % (92.0–97.3 %)	0.95 (0.92–0.97)

AUROC, area under the receiver operating characteristics curve.

Performance of trained AI on validation set

Overall accuracy of AI for prediction of neoplasia was 91.0 % (95 % CI: 89.1–92.7 %), with 97.3 % sensitivity (95 % CI: 95.4–98.5 %), 85.1 % specificity (95 % CI: 81.7–88.1 %), 85.9 % PPV (95 % CI: 82.7–88.7 %), 97.1 % NPV (95 % CI: 95.1–98.4 %) and 0.92 AUROC (95 %CI: 0.89–0.93). The AUROC curve for AI prediction in the body was significantly better than in the antrum (0.95 vs 0.90, $P=0.01$) and the corresponding accuracy of AI in the body was also better than in the antrum (0.95 vs 0.90, $P=0.01$). In terms of morphology, AI had statistically higher accuracy (98.2 % vs 91.4 % and 83.6 %, $P<0.05$) and AUROC (0.99 vs 0.92 and 0.91, $P<0.05$) in analyzing IIc lesions than IIa and IIb lesions (▶Table 2). Overall, AI is more confident in prediction of non-neoplastic than neoplastic lesions (84.5 % vs 81.8 %, $P<0.01$).

Validation set tesults

Performance of AI and the six junior endoscopists on the first part of the validation set is summarized in ▶Table 3. AI was better than all six endoscopists in accuracy (all $P<0.01$) and AUROC (all $P<0.01$). AI was also superior to individual endoscopists in sensitivity (AI vs II, III and IV; all $P<0.01$), specificity (AI vs I, III, V and VI; all $P<0.01$), PPV (AI vs I and VI; all $P<0.01$) and NPV (AI vs II, III, IV, VI; all $P<0.01$).

After revealing the AI prediction results from the first part of validation set to Group A endoscopists, their performance in the second part was summarized in ▶Table 4. In the second part, AI was still superior to all six endoscopists in accuracy (all $P<0.01$) and AUROC (all $P<0.01$). Specifically, AI was superior to individual endoscopists in terms of sensitivity (AI vs II and III, IV, V and VI; all $P<0.01$), specificity (AI vs I, V and VI; all $P<0.01$), PPV (AI vs I, V and VI; $P<0.01$), and NPV (AI vs II, III, IV, V and VI; all $P<0.01$).

The performance of the Group A endoscopists, to whom the AI prediction results from the first part of the validation set had been revealed, significantly improved in accuracy on the second part of the validation set (69.3 % to 74.7 %; $P=0.003$), AUROC (0.69 to 0.75, $P=0.018$), sensitivity (72.0 % to 82.7 %, $P=0.049$) and NPV (74.7 % to 82.5 % $P=0.003$). However, Group B endoscopists, who were unaware of the AI findings, significantly improved oinly in specificity (61.6 % to 70.4, $P<0.001$) but worsened in sensitivity (75.1 % to 64.6 % $P<0.001$) (▶Table 5). AI was better than the senior endoscopist in accuracy in the first part of the validation set (91.6 % vs 84.4 %, $P<0.01$) and AUROC (0.92 vs 0.84, $P<0.01$), but not in the second part of the validation set.

Discussion

We have developed an AI image classifier for characterization of gastric neoplastic lesions that is based on non-optical magnified endoscopic images obtained by NBI. The trained AI could achieve accuracy of >90 % and sensitivity of >97 % in predicting presence of neoplastic lesions, which was superior to all six junior endoscopists. Through feedback with AI prediction results, junior endoscopists showed significant improvement in predicting presence of neoplasia in gastric lesions in the second part of the validation study. In contrast, those who did not receive feedback from AI showed no improvement in accuracy of prediction and even worsened in sensitivity, further suggesting that AI feedback may shorten the learning curve for prediction of histology. In contrast, experienced endoscopist seemed to catch up quickly in the second part of the validation set in achieving performance comparable to the AI prediction.

Unlike most endoscopy centers in the rest of the world, those in Japan have ample experience in characterizing gastric neoplastic lesions. With the availability of trained AI, instant prediction of gastric lesion histology may be possible. More importantly, AI could also help to shorten the learning curve for less experienced endoscopists by providing immediate feedback like a virtual supervisor. Although there were initial concerns about the dependency of AI technology leading to deterioration of learned skills [16, 17], our study findings may suggest the opposite.

Traditionally, presence of a neoplastic lesion can be predicted by magnifying endoscopy with presence of a demarcation line together with irregular microvascular (MV) and microsurface (MS) pattern [4, 18]. With increasing use of high-definition endoscopic imaging, high-quality images can also be achieved with a non-magnifying endoscopy series by changing the depth of field of observation (e. g. near focus function), which can mimic the traditional optical magnifying image [19]. Use of NBI endoscopic images also helps to characterize endoscopic lesions better than white light endoscopy by AI [14]. The AI image classifier has a distinct advantage in analyzing these images with high accuracy and it is not surprising to find that a trained AI can differentiate the histology of gastric lesions better than trainee endoscopists. In fact, previous studies showed that the performance of AI was comparable to that of experts but did not exceed it [20, 21].

Lui Thomas K.L. et al. Feedback from artificial… Endoscopy International Open 2020; 08: E139–E146

E143

▶ **Table 3** Summary of the performance of AI and all endoscopists: first part of validation.

	Endoscopist							
	AI	Senior	I	II	III	IV	V	VI
Sensitivity	96.0% (93.4–98.6%)	88.1% (83.7%–91.6%)	96.0% (93.4–98.6%)	42.3% (35.8–48.8%)	77.1% (71.6–82.6%)	52.5% (45.9–59.0%)	87.9% (83.6–92.1%)	85.2% (80.5–89.8%)
Specificity	88.1% (84.3–91.9%)	79.8% (73.9–84.8%)	48.0% (42.1–54.0%)	94.2% (91.4–96.9%)	58.8% (53.1–64.6%)	82.7% (78.2–87.1%)	61.7% (56.0–67.4%)	40.4% (34.6–46.2%)
PPV	86.6% 82.4%–90.9%)	84.4% (79.7%–88.4%)	59.8% (54.7–64.8%)	85.4% (78.9–92.0%)	60.1% (54.5–65.8%)	70.9% (64.0–77.8%)	64.9% (59.5–70.2%)	53.5% (48.3–58.7%)
NPV	96.4% (94.2–98.7%)	84.4% (79.7–88.4%)	93.7% (89.7–97.7%)	67.0% (62.3–71.7%)	76.1% (70.5–81.9%)	68.4% (63.4–73.3%)	86.4% (81.6–91.1%)	77.2% (70.4–84.1%)
Accuracy[1]	91.6% (89.1–94.0%)	84.4% (80.9–87.5)	69.4% (65.3–73.4%)	71.1% (67.1–75.1%)	67.0% (62.9–71.1%)	69.2% (65.2–73.3%)	73.4% (69.5–77.2%)	60.4% (56.1–64.7%)
AUROC[1]	0.92 (0.89–0.95)	0.84 (0.81–0.87)	0.72 (0.68–0.77)	0.68 (0.63–0.73)	0.68 (0.63–0.73)	0.68 (0.63–0.72)	0.75 (0.71–0.79)	0.63 (0.58–0.68)
Mean confidence	84.0% (82.6–85.4%)	94.6% (60.0–100.0%)	92.5% (91.1–93.9%)	75.4% (74.5–76.2%)	75.0% (74.0–75.9%)	85.6% (84.6–86.7%)	87.1% (86.0–88.4%)	75.5% (74.5–76.5%)

PPV, positive predictive value; NPV, negative predictive value; AUROC, area under the receiver operating characteristics curve.
[1] AI is superior to all junior endoscopists in terms of accuracy and AUROC (all P<0.01). Number in brackets refer to 95 % confidence intervals

▶ **Table 4** Summary of the performance of AI and all endoscopists: second part of validation.

	Endoscopist							
	AI	Senior	I	II	III	IV	V	VI
Sensitivity	98.4% (96.8–99.9%)	87.6% (84.4–90.4%)	99.6% (98.7–99.9%)	60.8% (53.6–67.2%)	73.9% (68.2–79.6%)	39.1% (32.8–45.4%)	80.8% (75.8–86.0%)	73.9% (68.2–79.6%)
Specificity	82.4% (77.6–87.1%)	73.4% (67.3–79.1%)	51.5% (45.5–57.4%)	85.2% (81.0–89.4%)	81.5% (76.9–86.1%)	96.3% (94.0–98.6%)	55.6% (49.6–61.5%)	59.3% (53.4–65.1%)
PPV	84.8% (80.7–89.0%)	81.5% (76.9–85.6%)	63.6% (58.6–68.6%)	77.8% (71.7–83.8%)	77.3% (71.7–82.8%)	90.0% (84.1–95.9%)	39.2% (33.8–44.7%)	60.7% (55.0–66.4%)
NPV	98.1%. (96.3–99.9%)	99.4% (96.7–99.9%)	99.3% (97.9–99.9%)	71.9% (67.0–76.8%)	78.6% (73.8–83.4%)	65.0% (60.3–69.7%)	77.3% (71.4–83.2%)	72.7% (66.8–78.6%)
Accuracy[1]	90.4% (87.8–93.0%)	87.6% (84.4–90.4%)	73.6% (69.7–77.5%)	74.0% (70.2–77.8%)	78.0% (74.4–81.6%)	70.0% (65.9–74.0%)	67.2% (63.1–71.3%)	66.0% (61.8–70.1%)
AUROC[1]	0.91 (0.88–0.93)	0.90 (0.88–0.93)	0.75 (0.71–0.80)	0.73 (0.69–0.78)	0.78 (0.73–0.82)	0.68 (0.63–0.73)	0.68 (0.64–0.73)	0.67 (0.65–0.70)
Mean Confidence	82.3% (80.6–84.0%)	94.9% (70.0–100.0%)	90.4% (89.7–91.0%)	75.6% (74.9–76.3%)	75.2% (74.5–75.9%)	78.1% (77.1–79.1%)	87.5% (86.4–88.6%)	75.3% (74.4–76.3%)

PPV, positive predictive value; NPV, negative predictive value; AUROC, area under the receiver operating characteristics curve.
[1] AI is superior to all junior endoscopists in terms of accuracy and AUROC (all P<0.01). Number in brackets refer to 95 % confidence intervals.

▶ **Table 5** Comparison of the performance of Group A and Group B endoscopists.

	Group A Endoscopists			Group B Endoscopists		
	1st part of validation set	2nd part of validation set	*P*	1st part of validation set	2nd part of validation set	*P*
Sensitivity	72.0% (68.7–75.4%)	82.7% (79.8–85.6%)	0.049	75.1% (71.9–78.5%)	64.6% (61.0%-68.2%)	<0.001
Specificity	67.0% (63.7–70.1%)	68.1% (64.8–71.4%)	0.80	61.6% (58.3–64.9%)	70.4% (67.2–73.5%)	<0.001
PPV	63.9% (60.5–67.3%)	68.4% (65.1–71.6%)	0.29	61.2% (57.9–64.5%)	65.0% (61.5–68.6%)	0.50
NPV	74.7% (71.6–77.9%)	82.5% (79.5–85.5%)	0.049	75.5% (72.2–78.8%)	70.0% (66.9–73.2%)	0.12
Accuracy	69.3% (67.0%–71.6%)	74.7% (72.5%–77.0%)	0.003	67.7% (65.3%-70.0%)	67.7% (65.3%-70.1%)	0.11
AUROC	0.69 (0.67–0.72)	0.75 (0.72–0.77)	0.02	0.68 (0.66–0.71)	0.67 (0.65–0.70)	0.12
Mean confidence	80.7% (80.1–81.3%)	80.4% (79.9–80.9%)	0.88	82.8% (82.1–83.5%)	80.3% (79.7–80.9%)	<0.001

PPV, positive predictive value; NPV, negative predictive value; AUROC, area under the operator characteristics curve.

Another important observation was that the AI had more confidence in prediction of non-neoplastic lesions than neoplastic lesions. For non-neoplastic lesions, the MS and MV patterns were usually regular and variations were usually minimal when compared to neoplastic lesions [18]. Therefore, AI is more confident in predicting non-neoplastic lesions.

Our trained AI, which was based on still endoscopic images, will be very useful in further development of real-time AI diagnosis of gastric lesions. Given the high NPV (>97%), a negative response from AI would favor simple biopsy rather than complete resection of lesions. Moreover, AI can also be very useful in selection of the site of biopsy of a lesion. Traditionally, multiple biopsies have to be taken on a lesion to minimize sampling error but AI can identify the exact biopsy site for the best diagnostic yield. Because our AI image classifier is based on images from the readily available non-magnifying endoscopy system, it can be easily incorporated into an existing system without need of major equipment change.

This study has limitations. First, it is retrospective and the lesions were not a consecutive series, which could suffer from selection bias, particularly in selection of training and validation endoscopic images. Our AI image classifier analyzed static images, which were usually taken by endoscopists experienced in image-enhanced endoscopy. Second, inexperienced endoscopist may have a sampling issue by not choosing the correct region of interest of the lesions for AI interpretation, which may result in lower accuracy. Hence, a prospective real-time study involving endoscopists with variable experience is needed to validate our findings. Third, the current study focused on characterization rather than detection of gastric lesions. Because early gastric lesions can be very subtle, an endoscopist still needs to be able to identify the lesion prior to application of AI. However, application of AI for suspected lesions would take less time obtaining multiple biopsies and may potentially increase detection of subtle lesions that might otherwise not be biopsied.

Conclusion

We have developed an accurate AI image classifier for prediction of histology of gastric lesions based on non-magnified endoscopic images. The trained AI is better than junior endoscopists for histological prediction and it can also help speed the learning curve of junior endoscopists inb histological characterization of gastric lesions.

Competing interests

None

References

[1] Fitzmaurice C, Allen C, Barber RM et al. Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: a systematic analysis for the global burden of disease study global burden. JAMA Oncol 2017; 3: 524–548

[2] Yokota T, Ishiyama S, Saito T et al. Lymph node metastasis as a significant prognostic factor in gastric cancer: A multiple logistic regression analysis. Scand J Gastroenterol 2004; 39: 380–384

[3] Zheng Z, Liu Y, Bu Z et al. Prognostic role of lymph node metastasis in early gastric cancer. Chinese J Cancer Res 2014; 26: 192–199

[4] Kaise M. Advanced endoscopic imaging for early gastric cancer. Green J (ed.) Best Pract Res Clin Gastroenterol 2015; 29: 575–587

Lui Thomas K.L. et al. Feedback from artificial … Endoscopy International Open 2020; 08: E139–E146

E145

[5] Wang L, Huang W, Du J et al. Diagnostic yield of the light blue crest sign in gastric intestinal metaplasia: A meta-analysis. PLoS One 2014; 9: e92874

[6] Dinis-Ribeiro M, DaCosta-Pereira A, Lopes C et al. Magnification chromoendoscopy for the diagnosis of gastric intestinal metaplasia and dysplasia. Gastrointest Endosc 2003; 57: 498–504

[7] Morales TG, Bhattacharyya A, Camargo E et al. Methylene blue staining for intestinal metaplasia of the gastric cardia with follow-up for dysplasia. Gastrointest Endosc 1998; 48: 26–32

[8] Yao K, Anagnostopoulos GK, Ragunath K. Magnifying endoscopy for diagnosing and delineating early gastric cancer. Endoscopy 2009; 41: 462–467

[9] Chai N-L, Ling-Hu E-Q, Morita Y et al. Magnifying endoscopy in upper gastroenterology for assessing lesions before completing endoscopic removal. World J Gastroenterol 2012; 18: 1295–1307

[10] Gotoda T, Ho K-Y, Soetikno R et al. Gastric ESD: current status and future directions of devices and training. Gastrointest Endosc Clin N Am 2014; 24: 213–233

[11] Maekawa A, Kato M, Nakamura T et al. Incidence of gastric adenocarcinoma among lesions diagnosed as low-grade adenoma/dysplasia on endoscopic biopsy: A multicenter, prospective, observational study. Dig Endosc 2018; 30: 228–235

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

[13] Zhu Y, Wang Q-C, Xu M-D et al. Application of convolutional neural network in the diagnosis of the invasion depth of gastric cancer based on conventional endoscopy. Gastrointest Endosc 2019; 89: 806–815.e1

[14] Lui T, Wong K, Mak L et al. Endoscopic prediction of deeply submucosal invasive carcinoma with use of artificial intelligence. Endosc Int Open 2019; 07: E514–E520

[15] Brambilla E, Travis WD, Colby TV et al. The new World Health Organization classification of lung tumours. Eur Respir J 2001; 18: 1059–1068

[16] Rahwan I, Cebrian M, Obradovich N et al. Machine behaviour. Nature 2019; 568: 477–486

[17] Coiera E, Kocaballi B, Halamaka J et al. The price of artificial intelligence. IMIA Yearb Med Informatics 2019; 1: 1–2

[18] Yao K, Anagnostopoulos GK, Ragunath K. Magnifying endoscopy for diagnosing and delineating early gastric cancer. Endoscopy 2009; 41: 462–467

[19] Goda K, Dobashi A, Yoshimura N et al. Dual-focus versus conventional magnification endoscopy for the diagnosis of superficial squamous neoplasms in the pharynx and esophagus: A randomized trial. Endoscopy 2016; 48: 321–329

[20] Byrne MF, Chapados N, Soudan F et al. Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. Gut 2019; 68: 94–100

[21] Mori Y, Kudo SE, Misawa M et al. Real-time use of artificial intelligence in identification of diminutive polyps during colonoscopy a prospective study. Ann Intern Med 2018; 169: 357–366

E146

Lui Thomas K.L. et al. Feedback from artificial… Endoscopy International Open 2020; 08: E139–E146