

Multi-scale Hybrid Vision Transformer for Learning Gastric Cancer Histology

Yujin Oh^{a,1}, Go Eun Bae^b, Kyung-Hee Kim^c, Min-Kyung Yeo^{b,*}, Jong Chul Ye^{a,*}

^aKim Jaechul Graduate School of Artificial Intelligence, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon, 34141, Republic of Korea

^bDepartment of Pathology, Chungnam National University School of Medicine, Chungnam National University Hospital, Munwha-ro 282, Daejeon, 35015, Republic of Korea

^cDepartment of Pathology, Chungnam National University School of Medicine, Chungnam National University Sejong Hospital, 20 Bodeum 7-Ro, Sejong, 30099, Republic of Korea

ARTICLE INFO

Keywords: Vision Transformer, Digital Pathology, Histology, Gastric Cancer, Endoscopic Screening

ABSTRACT

Gastric endoscopic screening is an effective way to decide appropriate gastric cancer (GC) treatment at an early stage, reducing GC-associated mortality rate. Although artificial intelligence (AI) has brought a great promise to assist pathologist to screen digitalized whole slide images, existing AI systems are limited in fine-grained cancer subclassifications and have little usability in planning cancer treatment. We propose a practical AI system that enables five subclassifications of GC pathology, which can be directly matched to general GC treatment guidance. The AI system is designed to efficiently differentiate multi-classes of GC through multi-scale self-attention mechanism using 2-stage hybrid Vision Transformer (ViT) networks, by mimicking the way how human pathologists understand histology. The AI system demonstrates reliable diagnostic performance by achieving class-average sensitivity of above 0.85 on a total of 1,212 slides from multicentric cohort. Furthermore, AI-assisted pathologists show significantly improved diagnostic sensitivity by 12% in addition to 18% reduced screening time compared to human pathologists. Our results demonstrate that AI-assisted gastric endoscopic screening has a great potential for providing presumptive pathologic opinion and appropriate cancer treatment of gastric cancer in practical clinical settings.

© 2022

1. Introduction

Gastric cancer (GC) is reported as the fourth most leading cause of cancer death worldwide (Sung et al., 2021). Endoscopic screening is an effective way to detect GC at an early stage, guiding patients to get appropriate treatment according to their cancer stage. Although gastric endoscopic screening and proper early treatment have reduced GC-associated mortality rate (Jun et al., 2017), the increasing number of daily endoscopic biopsy cases adds an additional diagnostic workload to limited clinical resources. Accordingly, emerging application of artificial intelligence (AI) in the field of digital pathology has brought a great opportunity to effectively reduce diagnostic overloads, by automatically classifying massive number of digitalized whole slide images (WSIs).

AI applications in the field of digital pathology have already achieved powerful diagnostic performance in prostate or breast

cancer screening (Jiang et al., 2020; Bulten et al., 2020; Benjordi et al., 2017). AI-assistance systems have also been developed for GC (Yoshida and Kiyuna, 2021; Song et al., 2020; Yoshida et al., 2018; Park et al., 2021; Iizuka et al., 2020; Jang et al., 2021). An ideal fine-grained classification criteria of GC is presented by World Health Organization (WHO) (Nagtegaal et al., 2020); however, traditional AI systems mostly focus on detecting malignancy over benign or diagnosing only three subclassifications, i.e., benign, adenoma and carcinoma, which cannot fully cover detailed GC subclassifications. Indeed, differentiating fine-grained cancer subclassifications can be challenging even for pathologists, since adjacent diagnostic classes have region-level morphological similarities. Histologic GC classification needs comprehensive understanding of cell-level to tissue-level morphological features. In particular, when to diagnose a case confused between two adjacent classes, pathologists review slides by switching magnification from low to high resolution to understand both global structural relationships and regional morphological features.

Most existing AI systems for classifying GC employ convolutional neural network (CNN) (Song et al., 2020; Yoshida

*Corresponding authors: Tel.: +82-42-280-7196; fax: +82-42-280-7189; e-mail: mkyeo83@cnu.ac.kr (Min-Kyun Yeo) and Tel.: +82-042-350-4320; fax: +82-42-350-4310; e-mail: jong.ye@kaist.ac.kr (Jong Chul Ye)

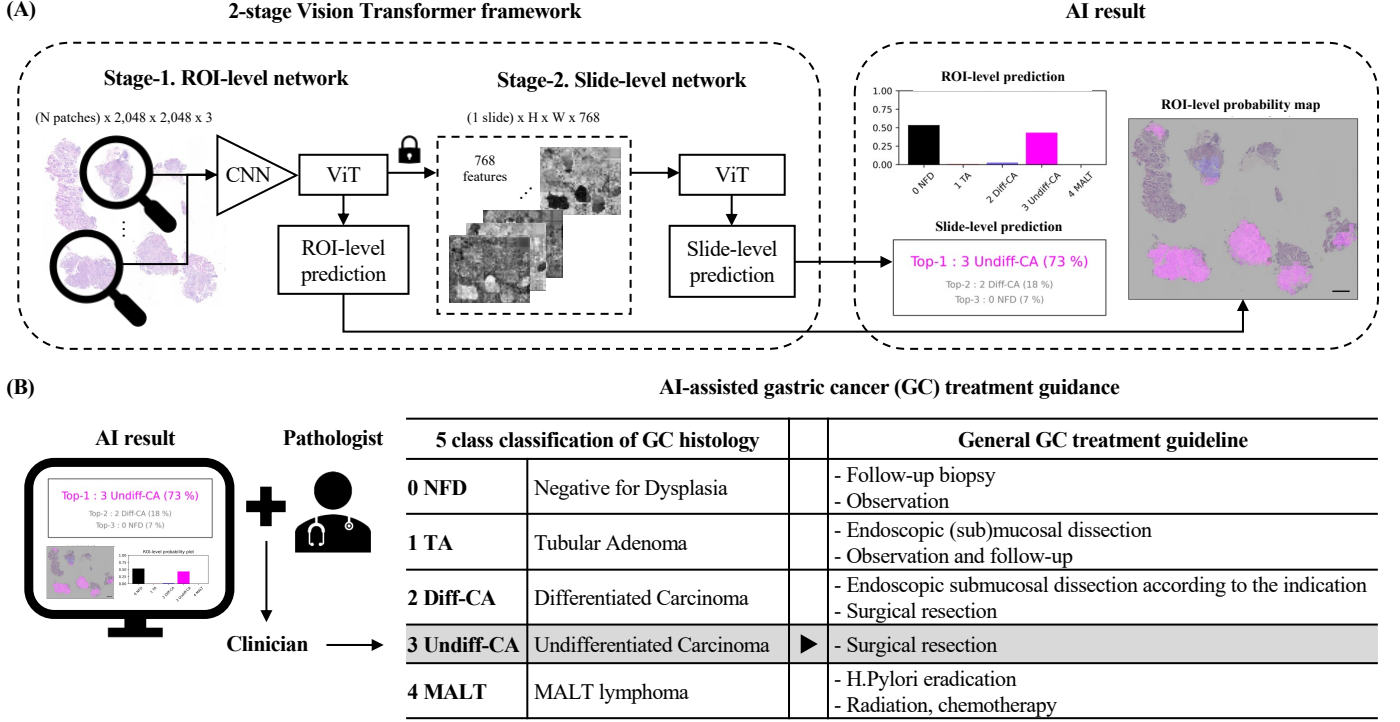


Fig. 1. Schematic of AI-assisted gastric cancer treatment guidance. (A) The AI system is composed of 2-stage Vision Transformer modules for comprehensive understanding of ROI-level and slide-level features. (B) Once a slide is diagnosed by the AI-assisted pathologist, the clinician can decide gastric cancer treatment based on the five subclassifications result.

et al., 2018; Park et al., 2021; Iizuka et al., 2020; Jang et al., 2021). In order to diagnose gigapixel-level WSIs using traditional CNN, patch-level training has become widespread by dividing WSIs into sub-patches (Liu et al., 2017). However, the patch-wise training is not effective in exploiting relationships between non-adjacent patches and understanding inter-patch structural relationships. To better understand global information, Park and colleagues proposed a RACNN, which aggregates patch-level features using additional convolution layers (Park et al., 2021). However, due to limited receptive field size of the convolutional kernel, the additional convolution layers still limit comprehensive understanding of non-adjacent inter-patch relationships.

In contrast, our contributions in this paper are multifold. First, inspired by recent success of Vision Transformer (ViT) (Dosovitskiy et al., 2020) that exploits long-range dependency between non-adjacent patches through multi-head self-attention mechanism, here we propose a patch-stacked hybrid ViT framework that can significantly expand receptive field. Specifically, as illustrated in Fig. 1(A), our system is composed of 2-stage ViTs: the first stage hybrid ViT encodes region-of-interest (ROI)-level histologic features from stacked patches, which is followed by the second stage ViT for understanding entire slide-level information. Specifically, multiple patch stacks from each WSI are fed into the ROI-level network to be trained to match their corresponding patch-stacked annotations. The ROI-level network inference results are then fed into the slide-level network and trained to diagnose its corresponding subclass. As a whole, we can expand receptive field up to entire size of WSI

for comprehensive prediction. In this way, the proposed AI system can mimic the entire process how pathologists understand WSIs.

Furthermore, our AI system can classify gastric endoscopic biopsies into 5 categories: negative for dysplasia (NFD), tubular adenoma (TA), differentiated carcinoma (Diff-CA), undifferentiated carcinoma (Undiff-CA), and MALT lymphoma (MALT), which is firstly tried in this study. As shown in Fig. 1(B), the proposed subclassifications can be matched to the general gastric cancer treatment guidance (jgca@koto.kpu-m.ac.jp, 2011), thus the results can be directly utilized to guide proper GC treatment in clinical settings.

Last but not least, the proposed AI system demonstrates promising GC classification performance by achieving average diagnostic sensitivity of above 0.85 for both internal and external cohort test set. Furthermore, in the observation study, pathologists assisted by the AI system shows significantly improved diagnostic performance by achieving average diagnostic sensitivity of 0.93 ± 0.06 compared to human pathologist performance of 0.83 ± 0.03 . The reliable performance of the AI system on multi-center daily cohorts and observer study demonstrates that the proposed AI system holds great promise in providing practical opinion for guiding appropriate treatment for early-stage GC patients.

2. Proposed framework

Vision Transformer (ViT) has shown state-of-the-art (SOTA) performance in various computer vision tasks (Khan et al.,

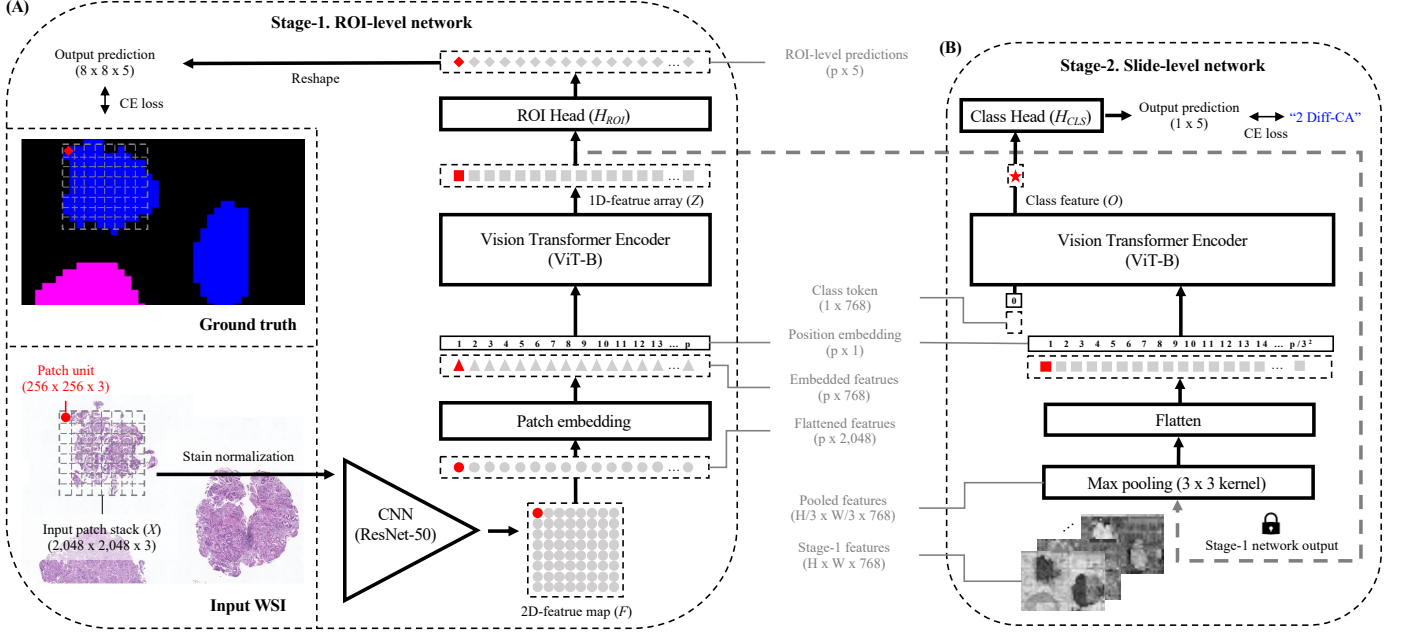


Fig. 2. Proposed 2-stage multi-scale hybrid Vision Transformer framework. (A) ROI-level prediction network with a hybrid Vision Transformer backbone. (B) Slide-level prediction network with a Vision Transformer backbone.

2021). By extending this idea, one of the important contributions of this paper is a 2-stage ViT architecture that effectively learns inter-patch relationships within gigapixel-level WSIs through multi-scale self-attention mechanism. As depicted in Fig. 2, the AI system is composed of ROI-level and slide-level networks. The ROI-level hybrid ViT network is composed of a convolutional neural network (CNN) module, a ViT encoder module and a ROI-level prediction head, which learn patch-level features from stacked patches. The slide-level ViT network is composed of a ViT encoder module and a slide-level classification head, which learn comprehensive information from stage-1 network output features over entire slide. More details are as follows.

2.1. Stage 1. ROI-level network

As illustrated in Fig. 2(A), for ROI-level prediction, a $2,048 \times 2,048$ -pixel input patch stack X is firstly stain normalized and fed into the CNN module. The CNN module extracts low-level features given the patch stack X , resulting a 8×8 grid 2-D feature map F :

$$F = \begin{bmatrix} f_1 & \cdots & f_w \\ \vdots & \ddots & \vdots \\ f_h & \cdots & f_p \end{bmatrix} = CNN \left(\begin{bmatrix} x_1 & \cdots & x_w \\ \vdots & \ddots & \vdots \\ x_h & \cdots & x_p \end{bmatrix} \right), \quad (1)$$

where f_p denotes a p -th feature map and x_p denotes a p -th input patch in the patch stack X .

The extracted features over the entire patch stack F are then flattened and embedded. The embedded patches added by positional embeddings are then fed into the ViT encoder module composed of successive self-attention layers to output a en-

coded feature array Z :

$$Z = [z_1, z_2, \dots, z_p], \quad (2)$$

where z_p denotes a p -th encoded feature in Z .

The encoded features Z are then linearly projected through the ROI-level prediction head and reshaped into its original 2D-feature map dimension. The ROI-level output predictions are trained to be matched with their corresponding patch-level ground truths by minimizing standard cross entropy (CE) loss:

$$\mathcal{L}_{CE_{stage-1}} = - \sum_{cls \in C} \sum_{p \in P} \mathbb{1}(y_p = cls) \log(H_{ROI}(z_p)), \quad (3)$$

where z_p denotes a p -th encoded feature in Z , H_{ROI} denotes the softmax probability of the ROI-level prediction head output, y_p denotes the ground truth label for a p -th patch, P denotes the set of patches in the patch stack, $\mathbb{1}(\cdot)$ denotes the indicator function, and C denotes the set of GC subclassification.

2.2. Stage 2. Slide-level network

As depicted in Fig. 2(B), the slide-level network takes intermediate features Z inferred by the stage-1 network over entire receptive field of an slide. To effectively cover the entire receptive field of the gigapixel-level WSI within limited patch grids of the ViT module, the stage-2 network adapted a max pooling module translating 3×3 adjacent patches into a representative 1×1 feature. As a whole, the entire slide-level receptive field can be expanded up to a 96×96 grid 2-D feature map, which correspond to a 12×12 mm slide.

The pooled input features are then flattened to be fed into the ViT encoder (ViT-B) module. In front of the flattened feature array, a learnable class token with identical feature dimension is prepended, which can attend to entire feature embeddings

throughout successive self-attention layers. The encoded class feature O is then linearly projected to be matched to its corresponding slide-level ground truth by minimizing class-weighted CE loss:

$$\mathcal{L}_{CE_{stage-2}} = - \sum_{cls \in C} w_{CLS} \mathbb{1}(y = cls) \log(H_{CLS}(O)), \quad (4)$$

where O denotes the encoded class feature, H_{CLS} denotes the softmax probability of the slide-level classification head output, y denotes the slide-level ground truth label, and w_{CLS} denotes a weight for adjusting unbalanced class distribution.

2.3. Network training

For training the ROI-level network, we utilized ResNet50 (He et al., 2016) and ViT-B (Dosovitskiy et al., 2020) for each CNN and ViT backbone, respectively. The batch size was set 4 and the model was trained for 30 epochs with the initial learning rate of 0.00001. For optimizing the ROI-level network training, we used Adam with decoupled weight decay (AdamW) optimizer with cosine scheduler for learning rate and weight decay. We applied geometric augmentation including random rotating, flipping and scaling with additional data augmentation techniques following BYOL (Grill et al., 2020) (color jittering, gaussian blurring and solarization).

For training the slide-level network, we utilized ViT-B (Dosovitskiy et al., 2020) as backbone. The batch size was set 5 and the model was trained for 50 epochs with initial learning rate of 0.0004. For optimizing the slide-level network training, we used the stochastic gradient descent (SGD) optimizer with cosine scheduler. For providing clinician-friendly results, the slide-level output was converted to probability score ranging from 0 to 100% by using Softmax function. In order to adjust false-positive rate gap between the training dataset and other test sets, we adapted a minimum cut-off threshold of 0.03 for the ROI-level probability. Specifically, a slide-level result with significantly low ROI-level probability was regarded as false-positive prediction. The optimal threshold for ROI-level offset were determined experimentally.

All the experiments were performed using Python version 3.9 and Pytorch library version 1.10 on a Nvidia RTX 3090 GPU.

3. Methods

3.1. Annotation of dataset

For our retrospective study, a total of 2,440 hematoxylin and eosin (H&E) stained slides were collected from the archives of Chungnam National University Hospital (CNUH) and Chungnam National University Sejong Hospital (CNUSH), and were scanned with a Panoramic 250 (3DHISTECH) scanner at $\times 40$ magnification. For the internal dataset, an expert pathologist annotated ROI-level labels into five subclasses using ImageJ (Abramoff et al., 2004). Cases with negative for any tumorous condition and background pixels were classified as NFD. Cases with low to high grade dysplasia were classified as TA. Malignant tumor was divided into Diff-CA, Undiff-CA, and MALT

lymphoma. Diff-CA and Undiff-CA were classified by following Japanese classification guideline that was suggested for endoscopic resection (jgca@koto.kpu-m.ac.jp, 2011). Diff-CA included well to moderately differentiated tubular/papillary adenocarcinoma, and Undiff-CA included poorly differentiated tubular/poorly cohesive/signet ring cell (SRC)/mucinous adenocarcinomas. MALT class was diagnosed followed by Wotherspoon criteria which was score 4 or above (Wotherspoon et al., 1993). For all the datasets, slide-level diagnosis was conducted based on ROI-level analysis by an expert pathologist. For confusing cases, consensus were made by more than 2 pathologists.

3.2. Datasets

The internal dataset were collected from CNUH, including 1,228 WSIs from endoscopic biopsy slides. The internal dataset were randomly split into train, validation and internal test set of 70, 10 and 20 percentile from entire dataset, respectively, following class distribution of entire dataset. A detailed dataset distribution is described in Table 1(A). For the convenience of training the AI system, one clear WSI per each slide was collected automatically and down-sampled at $\times 20$ magnification with resolution of $0.485 \mu m$ per pixel. $2,048 \times 2,048$ -pixel patch stacks were sampled from all the train set and maximum 100 patch stacks were collected from each WSI. Patch stacks containing foreground tissues below 30 percentile to the entire patch stack size were excluded from the train set.

The internal cohort test set was collected from CNUH, including 876 WSIs from endoscopic biopsy slides cohort from June 2021 to July 2021. A detailed dataset distribution is described in Table 1(B).

The external cohort test set was collected from CNUSH, including 336 WSIs from endoscopic biopsy slides cohort from September 2020 to February 2021. A detailed dataset distribution is described in Table 1(C).

3.3. Observer study design

The observer study was designed to evaluate the AI-assisted pathologist performance on daily gastric endoscopic screening. As shown in Fig. 3, from 117 cases of the sampled internal cohort test set, 2 observer test sets ($n = 25/\text{set}$) were prepared considering the class distribution and the diagnostic difficulty level. The diagnostic difficulty was stratified following previous trial (Song et al., 2020): level 1, easy to diagnose under low magnification; level 2, easy, but needed examination under high magnification; level 3, difficult, but ancillary test not required; level 4, challenging and ancillary test required. We excluded cases with the highest diagnostic difficulty, for avoiding potential diagnostic disagreement.

All the participants were given documented AI system performance and classification guidelines justified by an expert pathologist participated in the annotation process. For each test set, participants were given 25-minutes of time constraints. Neither case revision nor answer correction was restricted. Total 6 pathologists from CNUH and CNUSH were participated in the observer test. Among the participants, 4 pathologists of over 10 years of experience were equally split into different test groups.

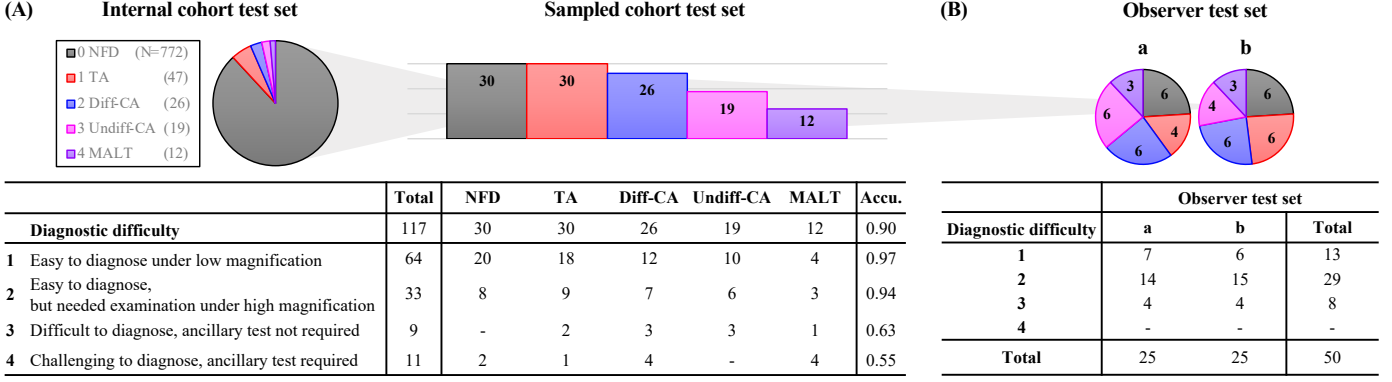


Fig. 3. Observer study design. (A) Diagnostic difficulty analysis. (B) Preparation of class distribution and diagnostic difficulty level-balanced observer test sets.

Table 1. Class distribution of training and test dataset.

(A) Internal dataset								
Class	Train set				Validation set		Test set	
	#slides	%	#patch stacks	%	#slides	%	#slides	%
0 NFD	263	31	1,850	15	38	28	73	30
1 TA	160	19	1,825	15	24	18	43	17
2 Diff-CA	175	21	2,977	24	31	23	54	22
3 Undiff-CA	180	21	5,032	40	33	24	58	23
4 MALT	68	8	826	7	9	7	19	8
Total	846		12,510	-	135		247	

(B) Internal cohort test set								
Class	Train set				Validation set		Test set	
	#slides	%	#patch stacks	%	#slides	%	#slides	%
0 NFD	-		-		-		772	88
1 TA	-		-		-		47	5
2 Diff-CA	-		-		-		26	3
3 Undiff-CA	-		-		-		19	2
4 MALT	-		-		-		12	1
Total	-		-		-		876	

(C) External cohort test set								
Class	Train set				Validation set		Test set	
	#slides	%	#patch stacks	%	#slides	%	#slides	%
0 NFD	-		-		-		297	88
1 TA	-		-		-		11	3
2 Diff-CA	-		-		-		12	4
3 Undiff-CA	-		-		-		8	2
4 MALT	-		-		-		8	2
Total	-		-		-		336	

3.4. Metrics

For evaluating the ROI-level network performance, we utilized patch-level accuracy per each WSI. The metric is defined as follows:

$$Accuracy_{patch}(y, \hat{y}) = \frac{1}{|S_c|} \sum_{s \in S_c} \left(\frac{1}{|P_s|} \sum_{p \in P_s} M_p \cdot 1(\hat{y}_p = y_p) \right),$$

where S_c is the set of WSI samples of each slide-level ground truth class, P_s is the set of patches within s -th WSI sample, \hat{y}_p is the predicted value of the p -th patch and y_p is the corresponding patch-level annotation label, and M_p is foreground mask, which is set to 1 when a patch consists of more than 10% of foreground pixels.

For evaluating the slide-level network performance in a class-wise manner, we utilized class-wise slide-level AUC (area under the ROC curve), specificity and sensitivity. To focus on multi-class classification capability of the AI system in clinical settings, we utilized diagnostic sensitivity as a primary criterion. All the metrics were by calculated using Scikit-learn package (Pedregosa et al., 2011).

For averaging multi-class metrics or multi-observer test metrics, we utilized macro average as default. The metric is defined as follows:

$$Average = \frac{1}{|C|} \sum_{cls \in C} M_{cls},$$

where C is the set of slide-level ground truth classes and M_{cls} represents any class-wise performance metric, e.g., AUC, specificity and sensitivity.

For averaging entire observer test records, e.g., confident level of the diagnosis and screening time, we utilized micro average. The metric is defined as follows:

$$Average(micro) = \frac{1}{|S|} \sum_{s \in S} s,$$

where S is the set of observer test records over all the participants.

A pathologist of over 5 years of experience and a pathology resident of over 4 years of experience were split into different test groups.

Participants were asked to diagnose WSIs in each observer test set with AI-assistance (AI-assisted) and without AI-assistance (Pathologist-only), respectively. The order of the AI-assistance and the pathologist-only trials was pre-determined, as described in Table 4. Minimum 3-hours of break time was given between each test sets. The participants were asked to fill out answers in tables (Excel 2019, Microsoft), and asked to score their confident level of the diagnosis from least (0) to most (1.00) for each case. Confident level of the diagnosis was justified as follows: indefinite for, <0.50; suspicious for, 0.51-0.70; favor, 0.71-0.80; consistent with, 0.81-0.90; diagnostic of, 0.91-1.00, following (Lindley et al., 2014).

3.5. Statistical analysis

Statistical analysis for the observer test was performed using MATLAB R2020a (Mathworks, Natick). Kolmogorov Smirnov test was used to evaluate normality of all the results. Since all the test results were non-normally distributed, a two-sided Wilcoxon rank sum test was used to statistically compare performance metrics. p-value was utilized as statistical significance level and indicated as asterisks, i.e., * for $p < 0.05$; ** for $p < 0.01$.

4. Experimental Results

4.1. ROI-level prediction performance

The proposed ROI-level network achieved patch-level accuracy of 0.85 ± 0.06 , as shown in Table 2(A). The class-wise diagnostic performance showed that our ROI-level network was mostly confused when predicting the Diff-CA class patches. Representative ROI-level probability maps were further analyzed in Appendix A.

Table 2. Diagnostic performance of the proposed AI system.

(A) Internal test				
Class	ROI-level <i>Accuracy_{patch}</i>	Slide-level AUC [95% CI]	Specificity	Sensitivity
0 NFD	0.97	0.99 [0.97, 1.00]	1.00	0.88
1 TA	0.82	0.99 [0.97, 1.00]	0.98	0.93
2 Diff-CA	0.79	1.00 [0.98, 1.00]	0.98	0.93
3 Undiff-CA	0.87	1.00 [0.98, 1.00]	0.96	1.00
4 MALT	0.83	1.00 [0.98, 1.00]	0.99	0.89
Average	0.85 ± 0.07	0.99 ± 0.00	0.98 ± 0.01	0.93 ± 0.05
(B) Internal cohort test				
Class	ROI-level <i>Accuracy_{patch}</i>	Slide-level AUC [95% CI]	Specificity	Sensitivity
0 NFD	-	0.96 [0.94, 0.97]	0.95	0.89
1 TA	-	0.97 [0.93, 1.00]	0.95	0.89
2 Diff-CA	-	0.98 [0.94, 1.00]	0.97	0.88
3 Undiff-CA	-	0.99 [0.97, 1.00]	0.97	1.00
4 MALT	-	0.96 [0.88, 1.00]	0.98	0.67
Average	-	0.97 ± 0.02	0.97 ± 0.02	0.87 ± 0.12
(C) External cohort test				
Class	ROI-level <i>Accuracy_{patch}</i>	Slide-level AUC [95% CI]	Specificity	Sensitivity
0 NFD	-	0.89 [0.85, 0.93]	0.97	0.80
1 TA	-	0.95 [0.87, 1.00]	0.95	0.73
2 Diff-CA	-	0.96 [0.88, 1.00]	0.94	0.75
3 Undiff-CA	-	0.97 [0.88, 1.00]	0.88	1.00
4 MALT	-	0.99 [0.94, 1.00]	0.96	1.00
Average	-	0.95 ± 0.04	0.96 ± 0.03	0.86 ± 0.13

4.2. Slide-level prediction performance

Slide-level prediction performance for each trial are provided in Table 2. The proposed slide-level network achieved the class-average sensitivity of 0.93 ± 0.05 , 0.87 ± 0.12 , and 0.86 ± 0.13 for each internal test, internal cohort test, and external cohort test, respectively. The comparative diagnostic sensitivity performance throughout three different trials demonstrated that the proposed model could be generalized to multicentre dataset.

We further counted error cases of each trial and analyzed representative falsely-classified WSIs in Appendix B. Error case analysis and Appendix C. False-positive case analysis.

4.3. Comparative network analysis

The proposed ROI-level performance was compared with several baseline networks, i.e., ResNet50, InceptionV3, and ViT-B, including traditional patch-level prediction methods utilized in comparative studies (Jang et al., 2021; Stegmüller et al., 2022). As shown in Table 3, the proposed patch-stacked hybrid ViT showed the most promising class-averaging patch-level accuracy compared to its counterparts with statistically significant level (p-value of < 0.001 for all the counterparts). Accordingly, as shown in Fig. 4, ROI-level probability map of the proposed patch-stacked hybrid ViT model provided the least confusing prediction compared to its counterparts.

The proposed slide-level performance was compared with several baseline methods, i.e., Tok-K Mean, Random Forest, and ResNet50, including traditional classification methods utilized in comparative studies (Song et al., 2020; Park et al., 2021). A combination of the proposed ViT as a stage-2 network along with the proposed stage-1 network showed the best class-averaging specificity and sensitivity among entire combinations of stage-1 and stage-2 networks. On the other hand, the stage-2 only network showed comparable or even inferior AUC performance to that of the Tok-K Mean method for both cohort test sets, since AUC measures binary classification performance of each sub-class. Therefore, the proposed 2-stage network demonstrated its superior multi-class classification performance to that of the traditional AI systems.

The proposed AI system performance was further evaluated between various experimental conditions, e.g., different magnifications and receptive field sizes of patch stacks. The proposed network showed best performance with magnification ratio of $\times 20$ magnification and ROI-level receptive field of 1×1 mm.

4.4. Observer test

The practical usability of the AI system was further evaluated on observation of AI-assisted pathologist performance. We observed that pathologist performance with AI-assistance exceeded that of pathologist-only trials for all metrics with statistically significant level, as shown in Table 4. The average specificity and sensitivity were increased by 0.02 and 0.10 with p-value of 0.015 and 0.006 respectively (Specificity from 0.96 ± 0.01 to 0.98 ± 0.01 and sensitivity from 0.83 ± 0.02 to 0.93 ± 0.06). We found that the improved performance was achieved under higher confident level of the diagnosis within shorter screening time, that the AI-assisted diagnostic confident level for each diagnosis improved by 0.04 with p-value of 0.007 (macro averaged confident level per WSI from 0.85 ± 0.13 to 0.89 ± 0.11). The AI-assisted trial saved screening time by 6.27 seconds per each WSI with p-value of 0.030 (micro averaged screening time per WSI from 34.80 ± 27.24 to 28.53 ± 23.15).

Moreover, the stand-alone AI performance (AI-alone trial) on both observer test sets was comparable to that of AI-assisted pathologist, and even exceeded pathologist-only trials by 0.12 and 0.03 for sensitivity and specificity, respectively.

We further evaluated receiver operating characteristics (ROC) curves over the entire internal cohort tests, as shown in Fig. 5. AI-assisted pathologist performance for all the

Table 3. Model performance comparison.

Stage-1 Network ROI-level	Magn.	Resolution ($\mu\text{m}/\text{pixel}$)	Receptive field ($\text{H} \times \text{W}$, mm)	Internal test set <i>Accuracy_{patch}</i>	Stage-2 Network	Internal test set			Internal cohort test set			External cohort test set		
					Slide-level	AUC	Spec.	Sens.	AUC	Spec.	Sens.	AUC	Spec.	Sens.
ResNet50	$\times 20$	0.49	0.13×0.13	0.79 ± 0.11	Top-K Mean	0.98	0.83	0.32	0.94	0.82	0.27	0.97	0.82	0.27
					Random Forest	0.98	0.97	0.90	0.88	0.89	0.71	0.89	0.89	0.58
					ResNet50	0.98	0.94	0.94	0.87	0.93	0.56	0.84	0.90	0.59
					ViT	0.99	0.98	0.94	0.95	0.96	0.74	0.93	0.94	0.83
InceptionV3	$\times 20$	0.49	0.13×0.13	0.80 ± 0.11	Top-K Mean	0.97	0.84	0.35	0.91	0.82	0.26	0.95	0.81	0.24
					Random Forest	0.97	0.97	0.88	0.84	0.88	0.60	0.90	0.89	0.71
					ResNet50	0.96	0.97	0.88	0.84	0.92	0.56	0.90	0.92	0.72
					ViT	0.98	0.98	0.91	0.95	0.9	0.70	0.94	0.92	0.76
ViT	$\times 20$	0.49	0.13×0.13	0.82 ± 0.10	Top-K Mean	0.98	0.86	0.45	0.94	0.83	0.28	0.96	0.84	0.36
					Random Forest	0.98	0.97	0.87	0.87	0.88	0.61	0.89	0.88	0.68
					ResNet50	0.97	0.98	0.92	0.80	0.91	0.44	0.88	0.89	0.61
					ViT	0.98	0.98	0.92	0.94	0.94	0.68	0.94	0.91	0.78
Patch-stacked hybrid ViT (Proposed)	$\times 20$	0.49	1×1	0.85 ± 0.06	Top-K Mean	0.99	0.88	0.49	0.97	0.87	0.41	0.98	0.86	0.47
					Random Forest	0.99	0.99	0.98	0.89	0.81	0.60	0.86	0.82	0.52
					ResNet	0.99	0.98	0.92	0.94	0.94	0.68	0.96	0.93	0.78
					ViT (Proposed)	1.00	0.99	0.97	0.97	0.97	0.87	0.95	0.96	0.86
	$\times 10$	0.97	1.5×1.5	0.85 ± 0.08	ViT	0.97	0.98	0.90	0.94	0.97	0.85	0.92	0.93	0.75
	$\times 5$	1.95	2×2	0.84 ± 0.07	ViT	0.92	0.93	0.75	0.94	0.95	0.82	0.91	0.93	0.80

Note: Magn.: magnification ratio; Spec.: Specificity; Sens.: Sensitivity

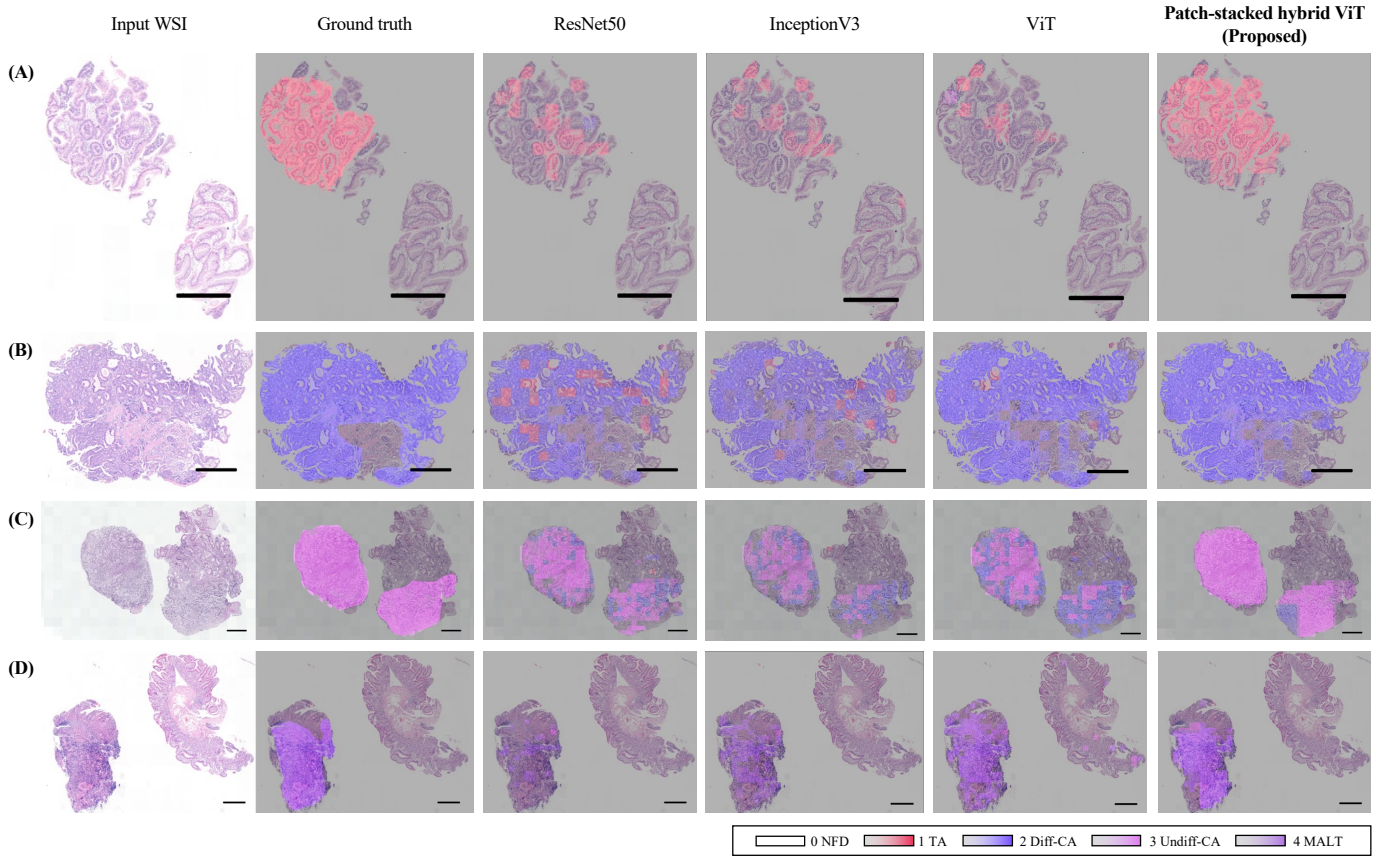


Fig. 4. ROI-level probability map comparison. Representative cases selected from (A) TA, (B) Diff-CA, (C) Undiff-CA and (D) MALT classes. All the scale bars indicate 500 μm .

classes were improved from Pathologist-only performance toward upper-left side of the curve near the AI-alone performance. We further discovered that the AI system helped pathologists make performance more consistent. The AI-assisted pathologist performance over all classes showed less variation compared to pathologist-only performance, as depicted as error bar of each trials.

5. Discussion

5.1. Reliable multi-class classification performance

The AI system achieves promising diagnostic accuracy on the the proposed five subclassifications. The AI system almost covers prevalent cases of daily endoscopic cohort except for diffuse large B-cell lymphoma, as shown in Table C.1. To the best of our knowledge, diagnosing 5 classes of gastric can-

Table 4. Observer test results.

Trial		Pathologists						Average (micro)	Improvement (w.r.t pathologist-only)	p-value
		1	2	3	4	5	6			
Pathologist-only	Observer test set	a	a	a	b	b	b			
	Specificity	0.97	0.95	0.96	0.95	0.96	0.95	0.96 ± 0.01		
	Sensitivity	0.85	0.83	0.83	0.80	0.85	0.80	0.83 ± 0.02		
	Confident level (a.u.)	0.76	0.78	0.87	0.85	0.89	0.96	(0.85 ± 0.13)		
	Screening time (s/WSI)	41.12	47.44	47.88	22.56	36.2	13.6	(34.80 ± 27.24)		
AI-assisted	Observer test set	b	b	b	a	a	a			
	Specificity	1.00	0.99	0.98	0.98	0.97	0.96	0.98 ± 0.01	+0.02	0.015
	Sensitivity	1.00	0.97	0.93	0.93	0.87	0.85	0.93 ± 0.06	+0.10	0.006
	Confident level (a.u.)	0.85	0.80	0.90	0.91	0.95	0.95	(0.89 ± 0.11)	+0.04	0.007
	Screening time (s/WSI)	20.56	34.52	44.24	20.16	35.84	15.88	(28.53 ± 23.15)	-6.27	0.030
AI-alone	Observer test set							a+b		
	Specificity							0.99	+0.03	
	Sensitivity							0.94	+0.12	

Note: a.u.: arbitrary unit

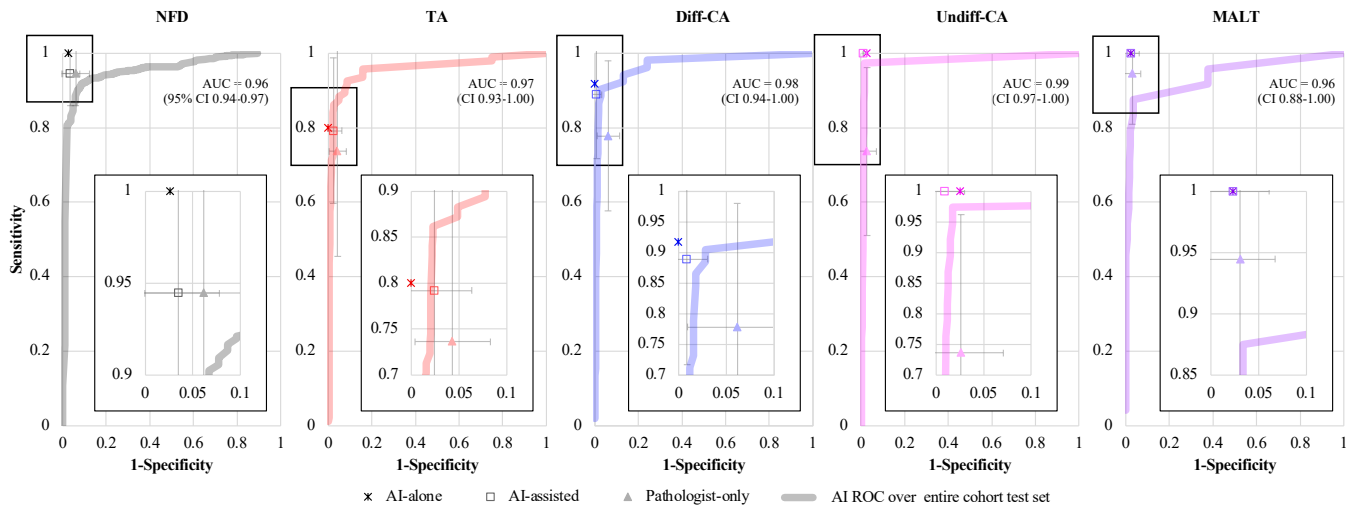


Fig. 5. Class-wise ROC curve and observer test results. ROC curves represent the proposed AI system performance on entire internal cohort test set and dot plots represent observer test performance of each AI-alone (asterisk), AI-assisted (square) and pathologist-only (triangle) trials.

cer, including MALT lymphoma, is firstly tried by the proposed AI system. Specifically, MALT lymphoma has a relatively high prevalence in East Asia related with high *Helicobacter pylori* (*H. pylori*) infection (Asenjo and Gisbert, 2007). In South Korea, 1-2% of patients receiving upper endoscopic biopsy are classified as MALT lymphoma, and it corresponds to 12% of patients who diagnosed as gastric malignancy (Yang et al., 2016). Diagnosis of MALT lymphoma largely depends on the pathologic confirmation, thus classification of MALT class should be listed on the gastric cancer screening program, especially in countries with high infection rates of *H. pylori*. Histologic features of MALT lymphoma help to diagnose the disease; however, diagnosis of MALT lymphoma is a challenge even for the pathologist and requires ancillary test, such as immunohistochemistry (IHC) or molecular evaluation, due to its morphologic similarities with other inflammatory or tumorous diseases (Bacon et al., 2007). The proposed AI system’s diagnostic sensitivity performance of MALT class (0.89, 0.67 and 1.00 for each internal test, internal cohort test and external cohort test set, respectively) demonstrates its capability for classifying MALT lymphoma and guiding clinician to proceed addi-

tional evaluation for the *H. pylori* infection.

The proposed classification additionally categories GC into differentiated and undifferentiated carcinoma. Undifferentiated-type carcinoma is reported to have high incidence of lymph node metastasis (Gotoda et al., 2000; Kook, 2019); thus, identifying undifferentiated carcinoma is important for deciding surgical treatment (Gotoda et al., 2006). A recent work reported sequential application of a differentiated/undifferentiated binary classifier on a normal/tumor classifier result (Jang et al., 2021). However, relatively poor performance of undifferentiated carcinoma class indicates the difficulty of discriminating confusing differentiated and undifferentiated cancer cells.

In fact, differentiating confusing adjacent classes, e.g., differentiated/undifferentiated and adenoma/adenocarcinoma, needs comprehensive understanding of structural characteristics. CNN-based patch-wise training has difficulty in understanding structural relationship between non-adjacent patches due to limited receptive field size. The proposed patch-stacked hybrid ViT models the long-range dependency among non-adjacent patch features after capturing short-range dependency among adjacent patch stacks exploited by CNN. As shown in

Fig. 4, the proposed patch-stacked hybrid ViT model provides stable ROI-level probability map with less confusing prediction between patches within tissues. Based on the improved ROI-level prediction results, the proposed slide-level prediction demonstrates average diagnostic sensitivity of above 0.85 for differentiating five subclassifications of GC.

5.2. Limitations

Our AI system is not free of limitations. As shown in Fig. 5, the AI system ROC performance for TA, Diff-CA and Undiff-CA exceed average performance of human pathologists. However, for NFD and MALT, average human pathologist performance exceeds the AI system ROC performance. One reasonable solution for improving degraded class performance may be matching class-wise distribution between training data set and daily cohort test set. For example, NFD distribution of the training patch stacks (15%) showed significant gap from that of the daily-acquired internal cohort test set (88%), as shown in Table 1. Similarly, daily internal cohort test sets contain around 30% of MALT cases that require ancillary test, as analyzed in Fig. 3. If our AI system would be trained with large number of NFD samples or samples required ancillary test, it may have predicted samples of daily endoscopic screening dataset with reliable performance.

Second, inter-centre stain variation may cause increased false-positive rate (18%) of the external cohort test compared to that of the internal cohort test (10%), as analyzed in Appendix C. This generalizability issue may be alleviated by developing advanced stain normalization techniques or training the system with multi-centre dataset.

5.3. Practical usability in clinical settings

Despite aforementioned limitations, our AI system has a strong point in providing explainable ROI-level probability map together with the slide-level prediction. As shown in Fig. A.1, pathologists can refer ROI-level probability maps to better understand the slide-level predictions, especially when the ROI-level probability distribution is non-dominant or the slide level probability is uncertain. Moreover, for region marked as suspicious for MALT class, as shown in Fig. B.2(B), pathologist can proceed ancillary test to confirm the final diagnosis of MALT lymphoma.

Finally, it should be noted that the results of the observer test show the stand-alone AI performance exceeds the average diagnostic performance of human pathologists for all classes. Therefore, integrating the AI system within the diagnostic workflow would be benefit of decreasing workloads on pathologists, while providing practical aid for planning surgical treatment, and can also provide diagnosis services for regions that have shortages in access to pathologists.

6. Conclusions

Previous AI assistance systems were developed as a pre-analytic tool for early attention of the suspicious lesion in the cases and give second opinion to the pathologists. We proposed

a 2-stage hybrid ViT-based AI system capable of reliable multi-class classification, which can be matched to general GC treatment guidance. Assisted by the proposed AI system, clinicians can receive a presumptive pathologic opinion for predicting prognosis and planning appropriate cancer treatment.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Research Foundation (NRF) of Korea grant NRF-2020R1 A2B5B03001980 and the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program (KAIST)). This work was also supported by Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HR20C0025).

Ethical approval

This study was approved by the institutional review board of Chungnam National University Hospital (CNUH) (IRB file no. 2021-10-028) and Chungnam National University Sejong Hospital (CNUSH) (IRB file no. 2021-12-004), which waived the requirement for informed consent. All slides were provided by the Biobank of CNUH, a member of the Korea Biobank Network.

Authors' Contributions

Yujin Oh contributed in formal analysis, investigation, software, validation, visualisation and writing – original draft; Go Eun Bae and Kyung-Hee Kim contributed in data curation; Min-Kyung Yeo contributed in conceptualisation, data curation, funding acquisition, methodology, supervision and writing– review & editing; Jong Chul Ye contributed in conceptualisation, project administration, funding acquisition, resources, supervision and writing– review & editing.

Appendix A. ROI-level probability map analysis

The probability maps generated by the proposed model highly correspond to the ground truth annotations, as shown in Fig. A.1. A close analysis showed that our model was mostly confused when distinguishing between TA/Diff-CA, as shown in Fig. A.1(B) and Diff-CA/Undiff-CA, as shown in Fig. A.1(E), due to their patch-level morphological similarity.

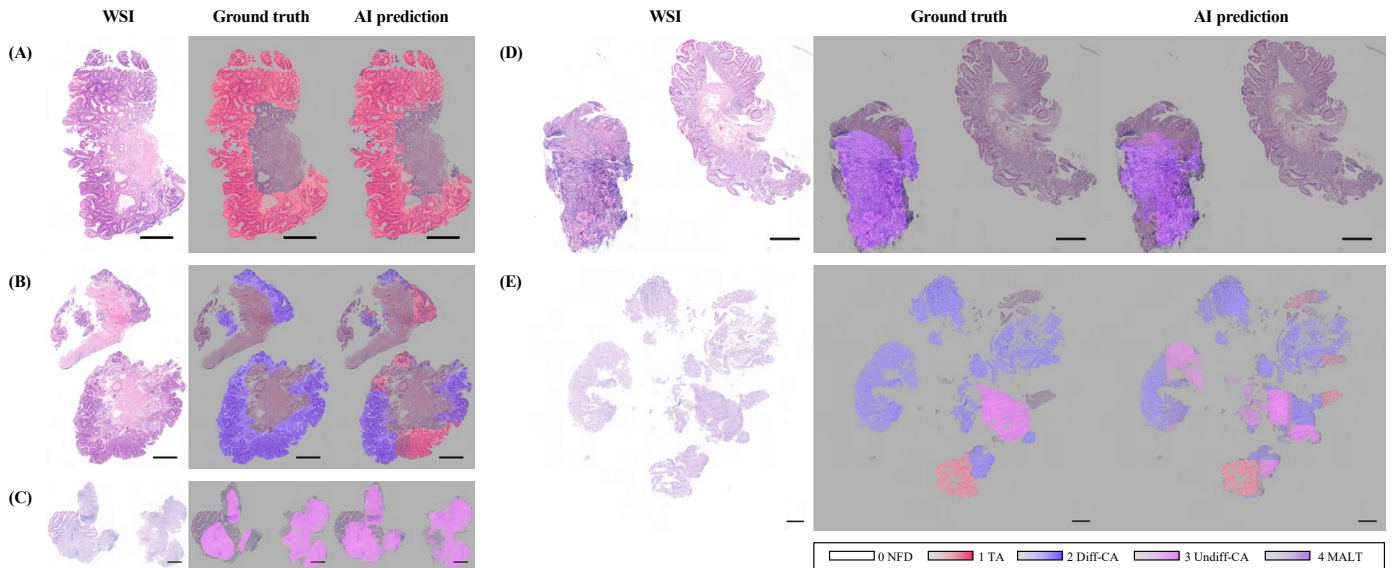


Fig. A.1. ROI-level probability map. (A) TA, (B) Diff-CA, (C) Undiff-CA, (D) MALT and (E) Undiff-CA. All the scale bars indicate 500 μm .

Appendix B. Error case analysis

We counted error cases of each test trial and analyzed representative error cases of each trial. For the internal test, most error cases were counted when distinguishing between adjacent classes (9 of 247 cases) and false-positive cases of NFD (9 of 73 cases). A representative result of WSI falsely diagnosed are shown in Fig. B.2(A). The representative result contains ROI-level probabilities of multiple classes, thus slide-level prediction is uncertain with probability score of under 90%.

For the internal cohort test, similar to the internal test results, most error cases were counted when distinguishing between adjacent classes (8 of 104 cases) and false-positive cases of NFD (87 of 772 cases). Errors cases were also counted when discriminating non-adjacent classes between NFD/MALT (3 of 12 cases) and TA/MALT (1 of 47 cases). Fig. B.2(B) depicts a representative case falsely diagnosed as negative for MALT class. The ROI-level probability map includes small portion of patches diagnosed as MALT class and under-estimated ROI-level probability. The results indicate that the AI system shows poor performance on slides that need ancillary test for accurate diagnosis.

For the external cohort test, all the error cases were counted when distinguishing between adjacent classes (6 of 39 cases) except for false-positive cases of NFD (59 of 297 cases), similar to the internal test results. A representative result of WSI falsely diagnosed are shown in Fig. B.2(C). The ROI-level probability for the falsely-classified class is below 0.1, thus slide-level prediction is uncertain with probability score of below 60%.

Appendix C. False-positive case analysis

In the internal test set and internal cohort test set, 4% (9 of 247 cases) and 10% (87 of 876 cases) of false-positive cases to entire dataset were observed, respectively. The falsely classified as non-NFD by detailed classification are summarized in Table

C.1. In the detailed NFD classes, hyperplastic polyp showed the highest false-positive rate of 19%, followed by acute and atrophic gastritis with 17% and 8%, respectively, whereas, 0% for foveolar hyperplasia.

In the external cohort test set, 18% (59 of 336 cases) of false-positive cases to entire dataset were observed. As summarized in Table C.1, acute gastritis showed the highest false-positive rate of 31%, followed by hyperplastic polyp and atrophic gastritis with 20% and 12%, respectively. Similar to the internal cohort test set, foveolar hyperplasia showed no false-positive case.

References

- Abràmoff, M.D., Magalhães, P.J., Ram, S.J., 2004. Image processing with imagej. *Biophotonics international* 11, 36–42.
- Asenjo, L., Gisbert, J., 2007. Prevalence of helicobacter pylori infection in gastric malt lymphoma: a sistematic review. *Revista Española de Enfermedades Digestivas* 99, 398.
- Bacon, C.M., Du, M.Q., Dogan, A., 2007. Mucosa-associated lymphoid tissue (malt) lymphoma: a practical guide for pathologists. *Journal of clinical pathology* 60, 361–372.
- Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermesen, M., Manson, Q.F., Balkenhol, M., et al., 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama* 318, 2199–2210.
- Bulten, W., Pinckaers, H., van Boven, H., Vink, R., de Bel, T., van Ginneken, B., van der Laak, J., Hulsbergen-van de Kaa, C., Litjens, G., 2020. Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study. *The Lancet Oncology* 21, 233–241.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Gotoda, T., Yamamoto, H., Soetikno, R.M., 2006. Endoscopic submucosal dissection of early gastric cancer. *Journal of gastroenterology* 41, 929–942.
- Gotoda, T., Yanagisawa, A., Sasako, M., Ono, H., Nakanishi, Y., Shimoda, T., Kato, Y., 2000. Incidence of lymph node metastasis from early gastric cancer: estimation with a large number of cases at two large centers. *Gastric cancer* 3, 219–225.

- Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al., 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Adv Neural Inf* 33, 21271–21284.
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: *CVPR*, pp. 770–778.
- Iizuka, O., Kanavati, F., Kato, K., Rambeau, M., Arihiro, K., Tsuneki, M., 2020. Deep learning models for histopathological classification of gastric and colonic epithelial tumours. *Scientific reports* 10, 1–11.
- Jang, H.J., Song, I.H., Lee, S.H., 2021. Deep learning for automatic subclassification of gastric carcinoma using whole-slide histopathology images. *Cancers* 13, 3811.
- Jiang, Y., Yang, M., Wang, S., Li, X., Sun, Y., 2020. Emerging role of deep learning-based artificial intelligence in tumor pathology. *Cancer Communications* 40, 154–166.
- jgca@koto.kpu-m.ac.jp, J.G.C.A., 2011. Japanese classification of gastric carcinoma: 3rd english edition. *Gastric cancer* 14, 101–112.
- Jun, J.K., Choi, K.S., Lee, H.Y., Suh, M., Park, B., Song, S.H., Jung, K.W., Lee, C.W., Choi, I.J., Park, E.C., et al., 2017. Effectiveness of the Korean national cancer screening program in reducing gastric cancer mortality. *Gastroenterology* 152, 1319–1328.
- Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M., 2021. Transformers in vision: A survey. *ACM Comput Surv* URL: <https://doi.org/10.1145/3505244>, doi:10.1145/3505244.
- Kook, M.C., 2019. Risk factors for lymph node metastasis in undifferentiated-type gastric carcinoma. *Clinical endoscopy* 52, 15.
- Lindley, S.W., Gillies, E.M., Hassell, L.A., 2014. Communicating diagnostic uncertainty in surgical pathology reports: disparities between sender and receiver. *Pathology-Research and Practice* 210, 628–633.
- Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G.E., Kohlberger, T., Boyko, A., Venugopalan, S., Timofeev, A., Nelson, P.Q., Corrado, G.S., et al., 2017. Detecting cancer metastases on gigapixel pathology images. *arXiv preprint arXiv:1703.02442*.
- Nagtegaal, I.D., Odze, R.D., Klimstra, D., Paradis, V., Rugge, M., Schirmacher, P., Washington, K.M., Carneiro, F., Cree, I.A., et al., 2020. The 2019 WHO classification of tumours of the digestive system. *Histopathology* 76, 182.
- Park, J., Jang, B.G., Kim, Y.W., Park, H., Kim, B.h., Kim, M.J., Ko, H., Gwak, J.M., Lee, E.J., Chung, Y.R., et al., 2021. A prospective validation and observer performance study of a deep learning algorithm for pathologic diagnosis of gastric tumors in endoscopic biopsies. *Clinical Cancer Research* 27, 719–728.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. *J Mach Learn Res* 12, 2825–2830.
- Song, Z., Zou, S., Zhou, W., Huang, Y., Shao, L., Yuan, J., Gou, X., Jin, W., Wang, Z., Chen, X., et al., 2020. Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning. *Nature communications* 11, 1–9.
- Stegmüller, T., Spahr, A., Bozorgtabar, B., Thiran, J.P., 2022. Scorenet: Learning non-uniform attention and augmentation for transformer-based histopathological image classification. *arXiv preprint arXiv:2202.07570*.
- Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F., 2021. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* 71, 209–249.
- Wotherspoon, A.C., Diss, T., Pan, L., Isaacson, P., Doglioni, C., Moschini, A., de Boni, M., 1993. Regression of primary low-grade B-cell gastric lymphoma of mucosa-associated lymphoid tissue type after eradication of *Helicobacter pylori*. *The Lancet* 342, 575–577.
- Yang, H.J., Lim, S.H., Lee, C., Choi, J.M., Yang, J.I., Chung, S.J., Choi, S.H., Im, J.P., Kim, S.G., Kim, J.S., 2016. Management of suspicious mucosa-associated lymphoid tissue lymphoma in gastric biopsy specimens obtained during screening endoscopy. *Journal of Korean Medical Science* 31, 1075–1081.
- Yoshida, H., Kiyuna, T., 2021. Requirements for implementation of artificial intelligence in the practice of gastrointestinal pathology. *World Journal of Gastroenterology* 27, 2818.
- Yoshida, H., Shimazu, T., Kiyuna, T., Marugame, A., Yamashita, Y., Cosatto, E., Taniguchi, H., Sekine, S., Ochiai, A., 2018. Automated histological classification of whole-slide images of gastric biopsy specimens. *Gastric cancer* 21, 249–257.

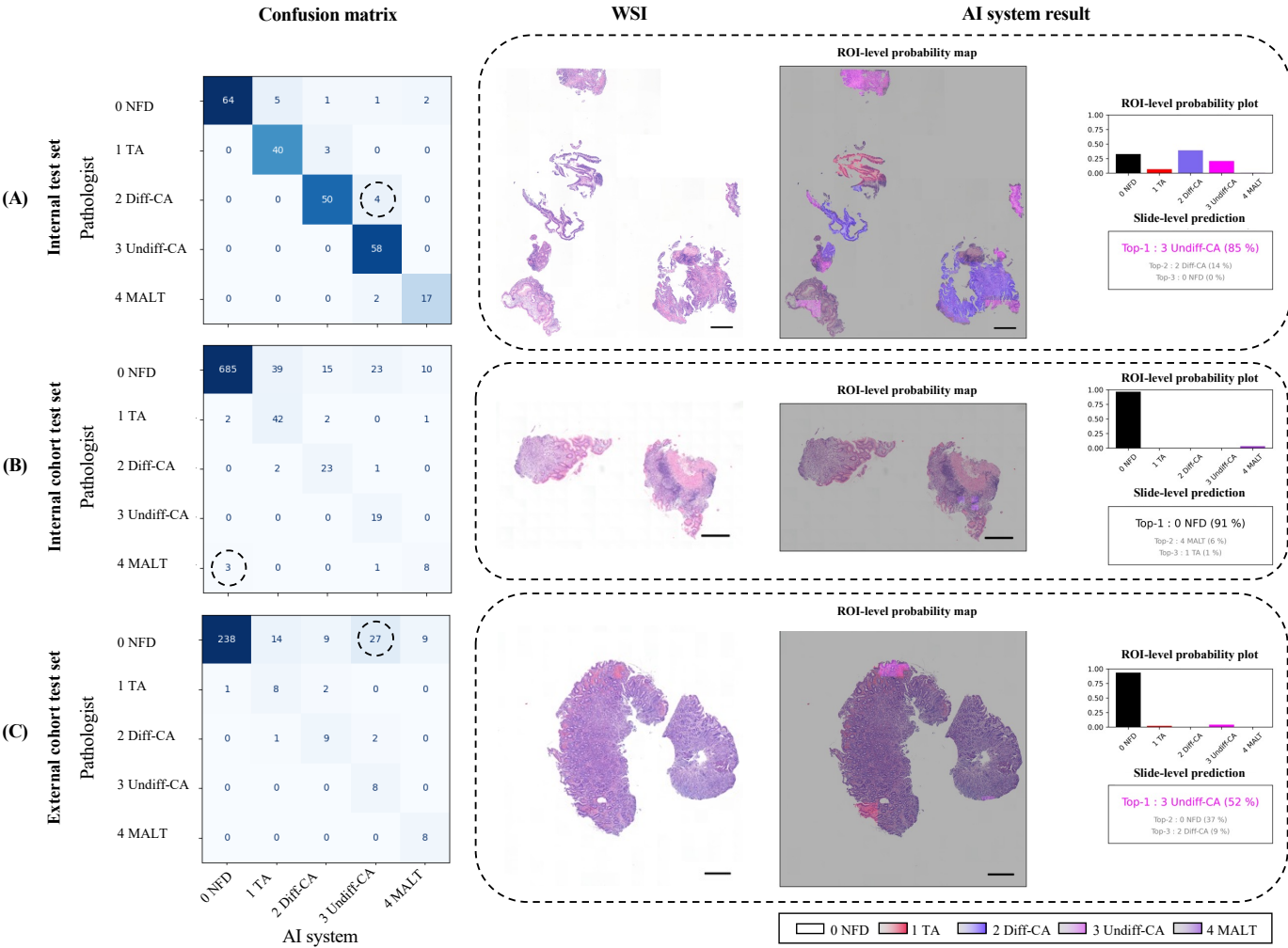


Fig. B.2. Representative error cases. Falsely-diagnosed cases (indicated as dashed lines) selected from (A) internal test set, (B) internal cohort test set and (C) external cohort test set. All the scale bars indicate 500 μm .

Table C.1. Detailed classification performance of the proposed AI system.

Detailed classification	(A) Internal cohort test set (N=876)						AI system prediction (per row percentile, %)				
	AI system prediction					Total					
	0 NFD	1 TA	2 Diff-CA	3 Undiff-CA	4 MALT		0 NFD	1 TA	2 Diff-CA	3 Undiff-CA	4 MALT
Negative for dysplasia											
- Foveolar hyperplasia	<u>10</u>					10	<u>100</u>				
- Hyperplastic polyp	<u>44</u>	3		7		54	<u>81</u>	6		13	
- Acute gastritis	<u>193</u>	11	9	10	9	232	<u>83</u>	5	4	4	4
- Atrophic gastritis	<u>438</u>	25	6	6	1	476	<u>92</u>	5	1	1	<1
Tubular Adenoma (TA)											
- TA, Low grade	2	<u>41</u>	1		1	45	4	<u>91</u>	2		2
- TA, High grade		<u>1</u>	1			2		<u>50</u>	50		
Differentiated carcinoma ; Tubular adenocarcinoma (TAC)											
- TAC, Well differentiated		2	<u>9</u>			11		18	<u>82</u>		
- TAC, Moderately differentiated			<u>14</u>	1		15			<u>93</u>	7	
Undifferentiated carcinoma											
- TAC, Poorly differentiated				<u>14</u>		14				<u>100</u>	
- Poorly cohesive carcinoma				<u>3</u>		3				<u>100</u>	
- Signet ring cell carcinoma				<u>2</u>		2				<u>100</u>	
MALT											
- MALT lymphoma	3			1	<u>8</u>	12	25			8	<u>67</u>
Total						876					
Detailed classification	(B) External cohort test set (N=337)						AI system prediction (per row percentile, %)				
	AI system prediction					Total					
	0 NFD	1 TA	2 Diff-CA	3 Undiff-CA	4 MALT		0 NFD	1 TA	2 Diff-CA	3 Undiff-CA	4 MALT
Negative for dysplasia											
- Foveolar hyperplasia	<u>4</u>					4	<u>100</u>				
- Hyperplastic polyp	<u>8</u>	1		1		10	<u>80</u>	10		10	
- Acute gastritis	<u>84</u>	6	6	17	9	122	<u>69</u>	5	5	14	7
- Atrophic gastritis	<u>142</u>	7	3	9		161	<u>88</u>	4	2	6	
Tubular Adenoma (TA)											
- TA, Low grade	1	<u>6</u>	1			8	13	<u>75</u>	13		
- TA, High grade		<u>2</u>	1			3		<u>67</u>	33		
Differentiated carcinoma ; Tubular adenocarcinoma (TAC)											
- TAC, Well differentiated			<u>2</u>			2			<u>100</u>		
- TAC, Moderately differentiated		1	<u>7</u>	2		10		10	<u>70</u>	20	
Undifferentiated carcinoma											
- TAC, Poorly differentiated				<u>5</u>		5				<u>100</u>	
- Poorly cohesive carcinoma				<u>3</u>		3				<u>100</u>	
- Signet ring cell carcinoma											
MALT											
- MALT lymphoma					<u>8</u>	8					<u>100</u>
Others (excluded)											
- Diffuse large B-cell lymphoma				1		1				100	
Total						337					