


RESEARCH ARTICLE



A deep learning approach for gastroscopic manifestation recognition based on Kyoto Gastritis Score

Ao Liu^{a†}, Xilin Zhang^{b,†}, Jiaxin Zhong^a, Zilu Wang^a, Zhenyang Ge^b , Zhong Wang^a , Xiaoya Fan^a  and Jing Zhang^b

^aSchool of Software Technology, Dalian University of Technology, Dalian, China; ^bDepartment of Digestive Endoscopy, Central Hospital of Dalian University of Technology, Dalian, China; ^cChina Medical University, Shenyang, China

ABSTRACT

Objective: The risk of gastric cancer can be predicted by gastroscopic manifestation recognition and the Kyoto Gastritis Score. This study aims to validate the applicability of AI approaches for recognizing gastroscopic manifestations according to the definition of Kyoto Gastritis Score, with the goal of improving early gastric cancer detection and reducing gastric cancer mortality.

Methods: In this retrospective study, 29013 gastric endoscopy images were collected and carefully annotated into five categories according to the Kyoto Gastritis Score, i.e. atrophy (A), diffuse redness (DR), enlarged folds (H), intestinal metaplasia (IM), and nodularity (N). As a multi-label recognition task, we propose a deep learning approach composed of five GAM-EfficientNet models, each performing a multiple classification to quantify gastroscopic manifestations, i.e. no presentation or the severity score 0–2. This approach was compared with endoscopists of varying years of experience in terms of accuracy, specificity, precision, recall, and F1 score.

Results: The approach demonstrated good performance in identifying the five manifestations of the Kyoto Gastritis Score, with an average accuracy, specificity, precision, recall, and F1 score of 78.70%, 91.92%, 80.23%, 78.70%, and 0.78, respectively. The average performance of five experienced endoscopists was 72.63%, 90.00%, 77.68%, 72.63%, and 0.73, while that of five less experienced endoscopists was 66.60%, 87.44%, 70.88%, 66.60%, and 0.66, respectively. The sample t-test indicates that the approach's average accuracy, specificity, precision, recall, and F1 score for identifying the five manifestations were significantly higher than those of less experienced endoscopists, experienced endoscopists, and all endoscopists on average ($p < 0.05$).

Conclusion: Our study demonstrates the potential of deep learning approaches on gastric manifestation recognition over junior, even senior endoscopists. Thus, the deep learning approach holds potential as an auxiliary tool, although prospective validation is still needed to assess its clinical applicability.

ARTICLE HISTORY

Received 24 April 2024
Revised 29 June 2024
Accepted 13 July 2024





KEYWORDS

Kyoto Gastritis Score; gastroscopic manifestations; gastric cancer; deep learning

1. Introduction

Gastric cancer (GC) is the fifth most common cancer globally and the fourth leading cause of cancer-related deaths [1]. In China, GC is the third most prevalent cancer and cause of cancer death, imposing a substantial social burden [2]. GC is often found in the late stage. Early diagnosis and treatment can greatly increase the five-year survival rate [3,4]. Endoscopic screening is crucial for early detection and treatment of GC [5–9]. Innovations like narrow-band imaging (NBI), magnifying endoscopy (ME), and chromoendoscopy have improved detection of precancerous lesions

linked to GC [10–14]. However, accurately diagnosing patients with early-stage GC and ensuring prompt, comprehensive treatment remains a major challenge in current medical practice. In China, only 20% of GC cases are detected early, significantly lower than Japan's 75% early detection rate [15]. This discrepancy is attributed to the country's vast population and limited endoscopists, hindering widespread endoscopic screenings. Regional healthcare discrepancies also lead to varied diagnostic practices among endoscopists, complicating consistent GC diagnoses. These challenges impede large-scale GC screenings.

CONTACT Jing Zhang  zj84402001@163.com  Department of Digestive Endoscopy, Central Hospital of Dalian University of Technology, Dalian, China; Xiaoya Fan  xiaoyafan@dlut.edu.cn  School of Software Technology, Dalian University of Technology, Dalian, China

[†]These authors contributed equally: Ao Liu, Xilin Zhang.

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

The progression of GC typically involves atrophy, intestinal metaplasia, dysplasia, and carcinoma with *Helicobacter pylori* (*H. pylori*) playing a key role in carcinogenesis [16,17]. Atrophy and intestinal metaplasia are precancerous states traditionally diagnosed through histopathology. Recent advancements in endoscopic technology now allow for the accurate identification of atrophic and metaplastic mucosa before histological confirmation. The Kyoto Global Consensus in 2013 introduced the Kyoto Classification of Gastritis (KCG) to standardize endoscopic diagnostic criteria, aligning with histopathological findings [18,19]. KCG is now a valuable tool that is widely accepted in clinical settings and global gastritis research [20,21]. The Kyoto Gastritis Score (KGS) is an extension of the KCG; it evaluates five manifestations related to increased risks of gastric cancer and *H. pylori* infection: atrophy, intestinal metaplasia, hypertrophic gastric fold, diffuse redness, and nodular gastritis [22]. Each manifestation is assessed for presence and severity, contributing to a total score that ranges from 0 to 8. A score of 0 suggests non-*H. pylori* infection, while scores of 2 or higher indicates its presence. A score of 4 or above is linked to a higher gastric cancer risk. Studies by Sugimoto et al. and Toyoshima et al. have further validated the association between the high KGS and gastric cancer risk, highlighting its potential in assessing cancer risk and managing of *H. pylori* infection [23,24].

Deep learning (DL) is a prominent subset of artificial intelligence (AI), known for its exceptional performance in image recognition, which has facilitated widespread exploration in the medical field. In gastro-intestinal endoscopy, AI has been utilized for tasks such as blind spot monitoring, lesion detection, and predicting the depth of malignant infiltration. However, its application in assessing gastric cancer risk is not extensively explored [16, 25–27].

To harness the full potential of AI in computer vision for effective gastric cancer risk assessment, this study developed a DL approach based on KGS, training a model on endoscopic images annotated with various KGS. The results demonstrated its ability to accurately identify and score the five endoscopic manifestations outlined in the KCG: atrophy, diffuse redness, enlarged folds, intestinal metaplasia, and nodularity. Comparative analyses also revealed that the AI's diagnostic capability surpassed that of both junior and senior endoscopists. Integrating AI into KCG can enhance the medical systems in assessing gastric cancer risk, addressing diagnostic variations among regions and endoscopists, reducing the workload of endoscopists, and potentially facilitating large-scale, targeted endoscopic screening.

2. Materials and methods

This study evaluated the artificial intelligence deep learning model in correctly identifying endoscopic manifestations and performing KGS on the gastric endoscopy images from visiting patients. The Flow chart of the experiment design is illustrated in Figure 1. Endoscopists assessed the images based on the KGS criteria and annotated the presence and severity of the five manifestations (Atrophy A, Diffuse Redness DR, Enlarged Folds H, Intestinal Metaplasia IM, and Nodularity N). These annotations were then reviewed and validated by auditing endoscopists to ensure accuracy and consistency.

Subsequently, the annotated images were processed and divided into training, validation, and testing sets. The training set was used to train five deep learning models, collectively named as DLKGS, with adjustments made to the model parameters based on the model performances on the validation set. This iterative process aimed to optimize DLKGS performance and construct an effective DL approach to KGS.

Finally, DLKGS performance was evaluated using the testing set. The accuracy and efficiency of DLKGS in identifying and scoring the endoscopic manifestations were compared with both highly experienced (senior) and less experienced (junior) endoscopists. This comparative analysis provided insights into the capabilities of DLKGS and its potential as a diagnostic tool in assisting endoscopists in KGS for gastric cancer risk assessment.

2.1. Study population

In total, we collected 29013 gastric endoscopy images from 2087 patients. Amongst the 2087 patients, 1134 (54.34%) were male patients and 953 were (45.66%) female patients, aging from 18 to 92 years with a mean age of 60.11 ± 12.07 years. All patients underwent upper gastrointestinal endoscopy examinations at the Digestive Endoscopy Center of the Central Hospital of Dalian University of Technology between July 2022 and November 2023. The inclusion criteria were as follows: (1) age over 18 years old; (2) voluntary participation in the clinical trial; (3) signing of the informed consent form that permits the use of their data for research purposes. The exclusion criteria were as follows: (1) use of medications within two weeks prior to enrollment that could affect the observations of this study, including proton pump inhibitors, H2 receptor antagonists; (2) history of gastric surgery, including surgical procedures, endoscopic submucosal

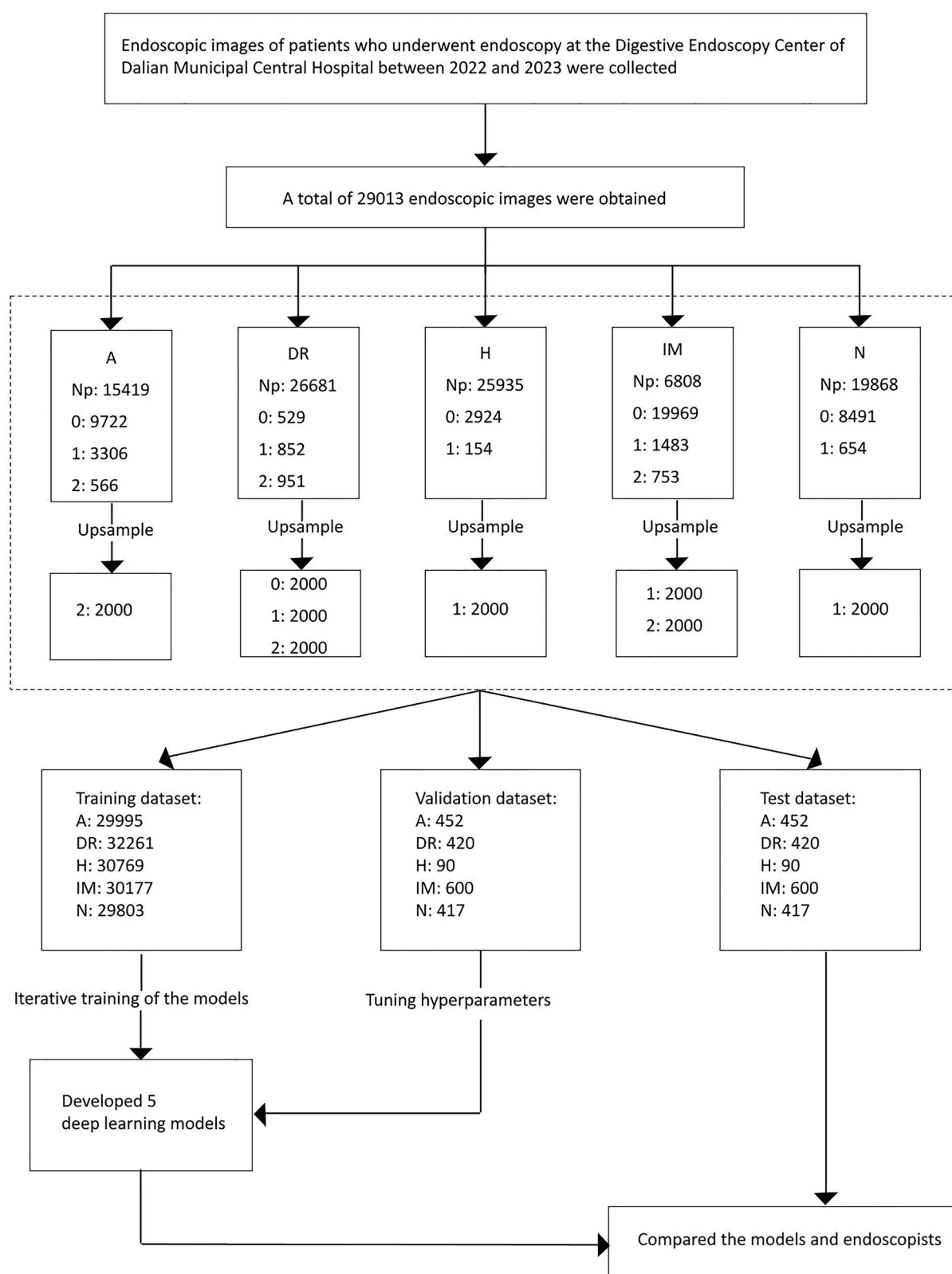


Figure 1. Flow chart of the experiment design.

dissection (ESD), or endoscopic mucosal resection (EMR); (3) severe cardiac, hepatic, renal dysfunction, as well as severe neurological or psychiatric

disorders. This study has been approved by the Ethics Committee of the Central Hospital of Dalian University of Technology (YN2022-047-06).

2.2. Collection of endoscopic image data

The eligible patients underwent upper gastrointestinal endoscopy examinations conducted by endoscopists, who performed comprehensive inspections and photography of the patients' stomachs with Olympus 260 and 290 (GIF-H260Z and EVIS LUCERA CV260/CLV260SL, GIF-H290Z and EVIS LUCERA ELITE CV290/CLV290SL). We obtained gastric endoscopy images from a selected number of patients by copying from the endoscopy device or the hospital's endoscopic reporting system. For rare endoscopic manifestations in patients (such as nodular gastritis N or severe atrophic gastritis A), gastric endoscopy videos were recorded using the Olympus endoscopy device and subsequently converted into sequences of images. The collected endoscopic images comprised two types: white light images (WLI) and narrow-band imaging (NBI) images.

2.3. Data preprocessing and annotation

We subjected the collected images to a screening process to ensure quality control. The exclusion criteria for the endoscopic image were defined as follows: (1) extracorporeal images; (2) images of the oropharynx, esophagus, or duodenum; (3) unclear images; (4) images severely obstructed by mucus, fluid, or other substances; (5) incomplete gastric emptying; (6) images of other types of lesions.

After the image screening, endoscopists with different experiences began the annotation process. The presence and severity of the five manifestations were annotated. It has to be noted that, NBI images were only scored for IM, whereas for other manifestations, they were annotated as 'No presentation'. The annotation does not imply overall scores of the patient, rather scores reflected only on single endoscopic image, which is given based on both presence of that manifestation and the location of stomach the image shows. Three junior endoscopists with over one year of experience and more than 1000 cases of endoscopic procedures annotated the selected images. All three junior endoscopists annotated each of the selected images. A senior endoscopist with over five years of experience and more than 5000 cases of endoscopic procedures reviewed the annotations. During the

review process, if there was any disagreement among the annotations provided by the junior endoscopists, additional consultations were sought from two other senior endoscopists to avoid mistakes and ensure the highest standard of accuracy in the annotations. The annotation guidelines followed the standard set by the KGS system, and all endoscopists have been trained on KGS before commencing the annotations.

To enhance DL model's robustness and prevent it from unnecessarily focusing on background information, all irrelevant metadata such as camera settings were removed from the images to reduce noise. Each image was randomly cropped to 384×384 pixels and subjected to a random horizontal flip with a probability of 50%, introducing variability and aiding the model in learning to recognize patterns independent of orientation. Subsequently, the images were normalized by adjusting their pixel values to have a mean of 0.5 and a standard deviation of 0.5 across all three RGB channels, improving model training stability and performance.

Table 1 presents the distribution of KGS in the endoscopic image datasets, i.e. the number and proportions of each endoscopic manifestation of Kyoto Gastritis and those of their respective KGS (0, 1, 2, or no manifestation). It reveals the proportion of images with higher scores is relatively low, hinting at an imbalance in the dataset.

The distribution of patients providing images for different scores of each endoscopic manifestation is shown in Table 2. The table lists the number of patients contributing images for each score (a total of 18 labels) and their proportions among all patients. As gastric endoscopy images can be captured from various angles of the same patient, images with different scores for the same manifestation may stem from the same patient. Consequently, this study divided the data on a per-image basis rather than on a per-patient basis. Table 2 highlights that the number of patients providing images of higher scores is relatively low.

The annotated images are partitioned as by the following; for each manifestation: (1) The images with the least occurrences were randomly split into training, validation, and testing sets with a 6:2:2 ratio. (2) Then, an equal number of images were randomly selected from the remaining images of different scores to match

Table 1. Gastric endoscopy images dataset for five gastroscopic manifestations.

	No presentation		Score 0		Score 1		Score 2		Sum
A	15419	(53.2%)	9722	(33.5%)	3306	(11.4%)	566	(2.0%)	29013
DR	26681	(92.0%)	529	(1.8%)	852	(2.9%)	951	(3.3%)	29013
H	25935	(89.4%)	2924	(10.1%)	154	(0.5%)	—		29013
IM	6808	(23.5%)	19969	(68.8%)	1483	(5.1%)	753	(2.6%)	29013
N	19868	(68.5%)	8491	(29.3%)	654	(2.3%)	—		29013

Table 2. Patient distribution in five gastroscopic manifestations.

	No presentation		Score 0		Score 1		Score 2	
A	1951	(93.5%)	1913	(91.7%)	967	(46.3%)	176	(0.8%)
DR	2058	(98.6%)	119	(5.7%)	294	(14.1%)	152	(0.7%)
H	2081	(99.7%)	946	(45.3%)	57	(0.3%)	—	
IM	1427	(68.4%)	2006	(96.1%)	316	(15.1%)	226	(10.8%)
N	2051	(98.3%)	1876	(89.9%)	12	(0.1%)	—	

Table 3. The test and validation image datasets for five gastroscopic manifestations.

	No presentation	Score 0	Score 1	Score 2	Sum
A	113	113	113	113	452
DR	105	105	105	105	420
H	30	30	30	—	90
IM	150	150	150	150	600
N	139	139	139	—	417

the number of images across all score categories within the validation or testing sets. (3) The remaining images with other scores were then assigned to the training set. The distribution of the validation and testing sets is shown in Table 3.

In the dataset of annotated images, a notable imbalance existed between images with low scores and those with high scores, resulting in a skewed sample distribution. To mitigate this imbalance and prevent the model from favoring information from more prevalent categories while neglecting rarer lesion types, this study employed the polynom-fit-SMOTE oversampling method to augment images with fewer occurrences (< 2000) to an appropriate quantity (2000) [28].

2.4. Model training

After evaluating popular models such as ResNet, ViT, and EfficientNet based on accuracy and runtime, we chose EfficientNet for this study due to its superior performance in both aspects [29–31]. We employed the GAM-EfficientNet [32] model to predict KGS from gastric endoscopy images. The GAM-EfficientNet model, derived from EfficientNetV2, incorporates a global attention mechanism (GAM) module, allowing the DL network to focus more on crucial information within the lesion area. EfficientNetV2 is well suited for clinical applications due to its fewer parameters and faster prediction speed. We trained five GAM-EfficientNet models specifically targeting each of the five manifestations of Kyoto gastritis, which we will collectively refer as DLKGS.

We used the PyTorch framework. All experiments were conducted on NVIDIA GeForce RTX 4090 devices for model training and testing. Transfer learning was applied using models pre-trained on ImageNet, with a learning rate of 0.0001, a decay coefficient of 0.01, and training epochs set to 200.

2.5. Model evaluation methods

We calculated common performance metrics for multi-class classification, i.e. average accuracy, specificity, precision, recall, and F1 score, as defined in Equations 1–5, to evaluate the overall classification performance. We also employed the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC). These metrics provide a comprehensive understanding of the approach's performance in accurately classifying different endoscopic manifestations. The ROC curve and AUC offer insights into the model's ability to differentiate between classes, while the average accuracy, specificity, precision, recall, and F1 score provide a more detailed evaluation of the classification performance across multiple classes.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

$$\text{Precision} = TP / (TP + FP) \quad (2)$$

$$\text{Specificity} = TN / (TN + FP) \quad (3)$$

$$\text{Sensitivity} = TP / (TP + FN) \quad (4)$$

$$\text{F1 score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (5)$$

TP, TN, FP, FN indicates true positive, true negative, false positive, and false negative, respectively.

2.6. Comparison of the diagnosis performance with endoscopists

In the evaluation of DLKGS with medical experts, we selected ten endoscopists from the Digestive Endoscopy Center of Central Hospital of Dalian University of Technology for evaluation. These endoscopists were divided into two groups based on their diagnostic experience: senior endoscopists with over 5 years of endoscopic work experience and junior endoscopists with over 1 year of work experience. After training on KGS, these endoscopists independently assessed the images in the testing set and assigned

scores according to the KGS guideline. The assessments of all endoscopists were then aggregated to analyze and compare their recognition abilities with DLKGS. To ensure an equal comparison between endoscopists and DLKGS, the metrics calculated were also the average accuracy, precision, specificity, recall, and F1 score for junior, senior, and all endoscopists.

To determine if there were significant differences in diagnostic abilities between DLKGS and senior/junior endoscopists, independent sample t-tests were computed using R. The significance level was set to 0.05. This statistical analysis aimed to provide insights into the comparative performance of DLKGS and the endoscopists of varying experience levels, shedding light on the effectiveness and reliability of the deep learning model in diagnosing Kyoto Gastritis manifestations.

3. Results

3.1. Model prediction

DLKGS demonstrates good overall classification performance on the test set for the five manifestations of gastric endoscopic images, with an average accuracy, specificity, precision, recall, and F1 score of 78.70%, 91.92%, 80.23%, 78.70%, and 0.78, respectively. Notably, DLKGS excels at scoring nodular gastritis (N), with an accuracy of 93.85%. The model also demonstrates superior average specificity, precision, recall, and F1 score for nodular gastritis (N), compared to the other manifestations, with values of 93.85%, 96.92%, 94.03%, 93.85%, and 0.94, respectively. However, the limited number of patients providing images with nodular gastritis (N) poses challenges to DLKGS generalization ability (Table 4).

3.2. Confusion matrix

The confusion matrix of the KGS models on the testing set are shown in Figure 2. Analysis of the matrix reveals that top-performing manifestation for DLKGS is nodular gastritis (N), followed by atrophic gastritis (A), characterized by relatively high recall rates. This implies that false negatives are less likely to occur in the

identification of these two manifestations. Despite the strong performance, nodular gastritis (N) faces challenges related to dataset limitations, which may impact DLKGS generalization ability. This imbalance in the dataset may impact DLKGS performance when applied in broader, practical settings, emphasizing the importance of addressing dataset limitations for robust and reliable model performance.

The confusion matrix analysis highlights less accurate recognition for atrophy score 2 (A2), diffuse redness score 1 (DR1), enlarged folds score 1 (H1), and intestinal metaplasia score 2 (IM2). This may stem from the scarcity of images in these categories in the dataset, leading to insufficient feature recognition by DLKGS. Additionally, some images may lack clear, visible features, further impacting the recognition of features. For example, the diagnosis of diffuse redness score 1 (DR1) relies on the partial visibility of the sub-epithelial capillary network upon close observation of the gastric mucosa. However, judging the surface condition accurately from a distance can be challenging, potentially leading to confusion with cases exhibiting no diffuse redness manifestation (DR no manifestation). Similarly, identifying enlarged folds score 1 (H1) requires observing gastric body folds with adequate insufflation, which could be misconstrued with images showing inadequately spread folds. In the case of intestinal metaplasia score 2 (IM2), the limited field of view resulting from close observation of the mucosa may complicate determining whether the anatomical site is the gastric body or the gastric antrum, causing confusion with cases of intestinal metaplasia score 1 (IM1) localized to the gastric antrum. These challenges underscore the importance of addressing dataset limitations and refining image acquisition protocols to enhance the model's accuracy and robustness in recognizing manifestations with intricate diagnostic criteria.

3.3. ROC and P-R curves

The performance of DLKGS was evaluated using ROC curves and Precision-Recall (P-R) curves. Figure 3 depicts the ROC curves and P-R curves, showcasing the model's proficiency in recognizing the five

Table 4. Diagnostic performance of DLKGS on the test set.

manifestations	Accuracy (%)	Specificity (%)	Precision (%)	Recall (%)	F1 score	Number of test set images
A	79.42	93.14	81.21	79.43	0.79	113
DR	75.24	91.75	76.22	75.24	0.75	105
H	76.67	88.33	78.03	76.67	0.77	30
IM	68.33	89.45	71.65	68.33	0.68	150
N	93.85	96.92	94.03	93.85	0.94	130
Overall	78.70	91.92	80.23	78.70	0.78	528

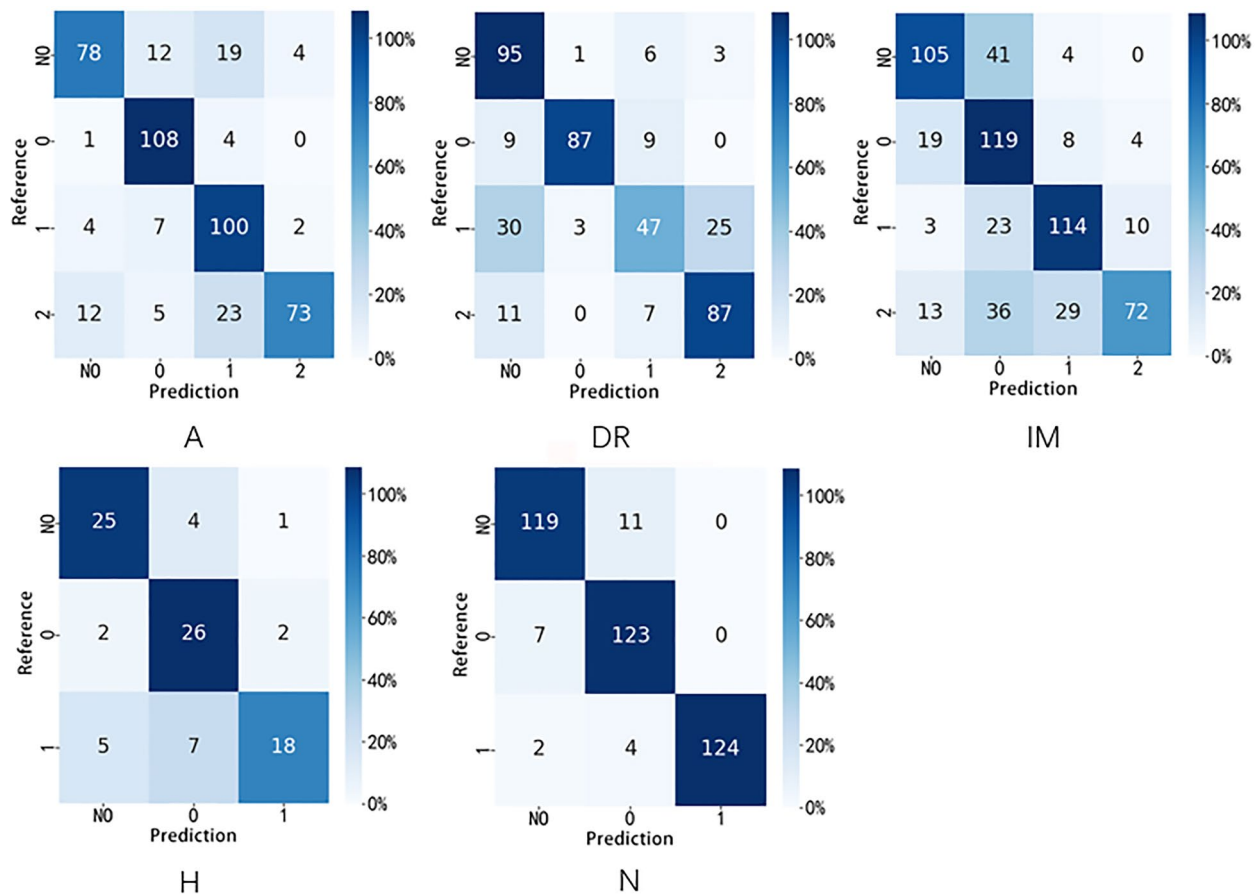


Figure 2. Confusion matrix of prediction in the test dataset for DLKGS.

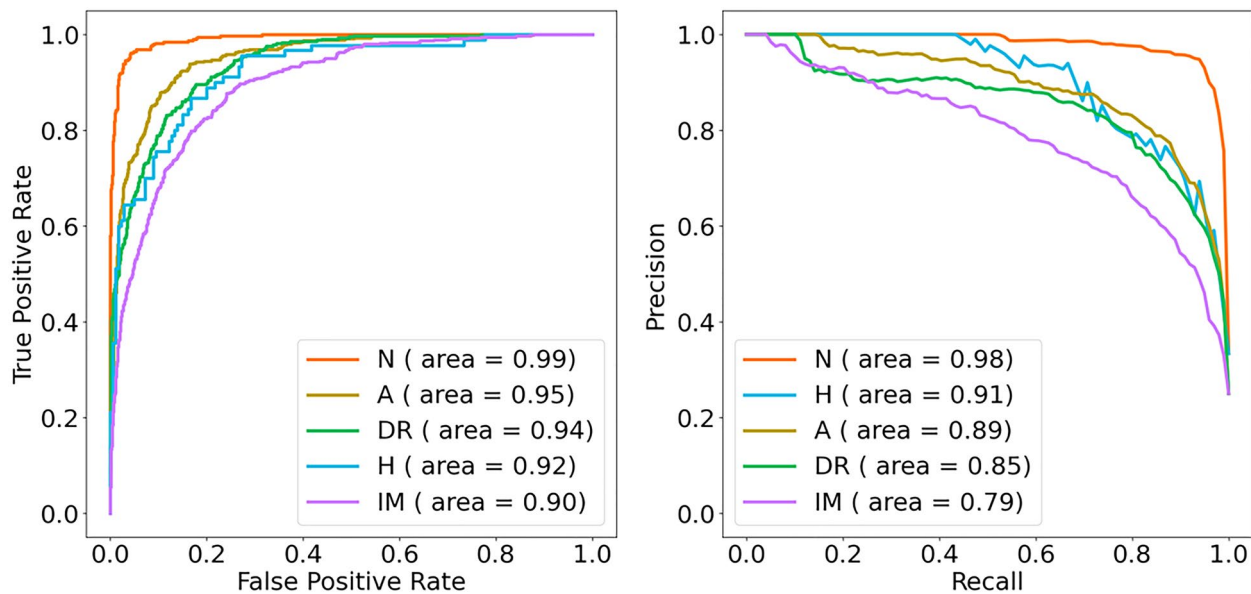


Figure 3. ROC and P-R curves of the five models.

manifestations of Kyoto gastritis on the testing set. In the testing set, the areas under the ROC curves for the recognition of A, DR, H, IM, and N manifestations were 0.95, 0.94, 0.92, 0.90, and 0.99, respectively, with

corresponding P-R values of 0.89, 0.85, 0.91, 0.79, and 0.98. The model exhibited the highest recognition performance for the N manifestation, followed by the A manifestation. The recognition abilities for DR and H

manifestations were comparable, while the recognition ability for IM manifestation was the lowest.

3.4. Visual explanations of models

We perform a visualization of DLKGS with Gradient-weighted Class Activation Mapping (Grad-CAM) [33]. The evaluation results of the five different backbone networks combined with attention modules are summarized in Figure 4. Atrophic gastritis score 1 (A1) includes cases where the atrophy boundary extends beyond the angulus but does not reach the cardia. We present only one case scenario for each manifestation across the possible scores for demonstration, except for IM, for which two cases, one NBI image and one WLI image are shown (the 4th and 5th row in Figure 4). The results indicate that the five models have achieved good performance after the addition of the attention module. The highlighted part (red area) in the figure is the area with a higher weight adopted by DLKGS when classifying endoscopic images. The model has been trained to classify and evaluate heavily based on the highlighted part. In contrast, the remaining

light-colored parts (blue parts) are the areas with lower weights adopted. Consequently, the attention weight distribution of DLKGS in classifying endoscopic images can be understood by observing the distribution of different colors in the heat map generated by Grad-CAM, providing a visual interpretation of the model classification results.

3.5. Assessment of endoscopist recognition abilities

The recognition abilities of all ten endoscopists for different manifestations on the testing set are presented in Table 5. The data indicates that both senior and junior endoscopists demonstrated excellent recognition abilities for nodular gastritis (N) in comparison to the other four manifestations. However, their proficiency in recognizing intestinal metaplasia (IM) was notably lower, which is consistent with the model's recognition abilities. Table 5. provides insights into the varying performance levels of the two groups of endoscopists across different manifestations, shedding light on areas where additional training or support

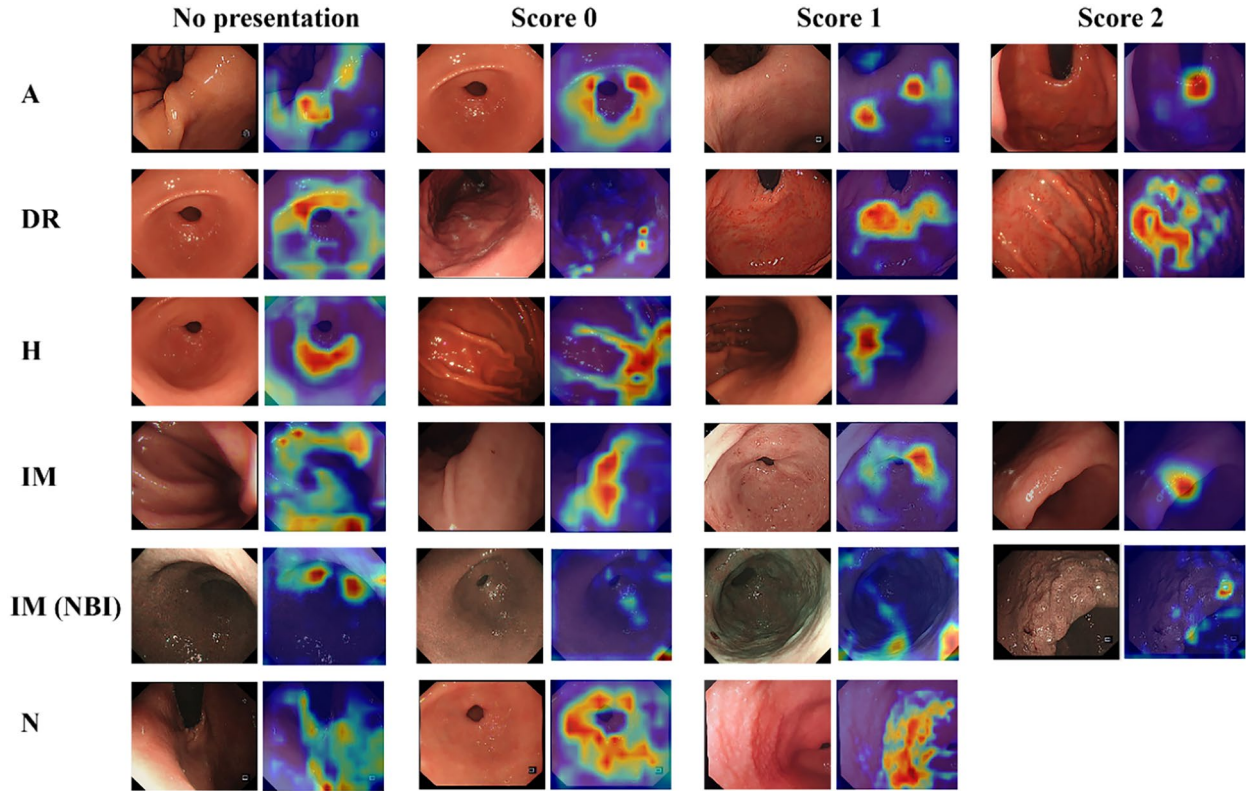


Figure 4. Representative examples of different manifestations with different scores (left) along with their feature heatmaps outputted by Grad-CAM for DLKGS (right). Since each image is annotated with 5 labels (each corresponding to one manifestation), identical images exist to represent different manifestations. For example, the image illustrating score 0 for manifestation A is also a representative image showing no presentation for manifestation DR. All representative images for A, DR, H, and N are WLI images, whereas for IM, both WLI images and NBI images are shown

may be beneficial to enhance diagnostic accuracy and consistency.

3.6. Comparison of model and endoscopist recognition abilities

The results depicted in Figure 5 demonstrate that the overall identification capability of DLKGS surpasses the average proficiency of all endoscopists. One-sided t-tests comparing the recognition abilities of DLKGS with both senior and junior endoscopists, as well as all endoscopists combined, are summarized in Table 6.

Table 5. Recognition ability of different gastroscopic manifestations in senior, junior, and mixed groups (* indicates $p < 0.05$).

	Accuracy (%)	Specificity (%)	Precision (%)	Recall (%)	F1 score
senior					
A	74.20	91.40	77.54	74.20	0.74
DR	64.71	88.24	70.09	64.71	0.66
H	82.00	91.00	82.70	82.00	0.82
IM	50.60	83.53	66.20	50.60	0.49
N	91.64	95.82	91.86	91.64	0.92
junior					
A	65.44	88.48	68.78	65.44	0.65
DR	60.24	86.75	66.62	60.24	0.62
H	70.44	85.22	71.13	70.44	0.70
IM	50.00	83.33	60.81	50.00	0.48
N	86.87	93.44	87.04	86.87	0.87
Mixed					
A	69.82	89.94	73.16	69.82	0.70
DR	62.48	87.50	68.36	62.48	0.64
H	76.22	88.11	76.92	76.22	0.76
IM	50.30	83.43	63.51	50.30	0.49
N	89.26	94.63	89.45	89.26	0.90

These tests reveal that the model consistently outperforms junior endoscopists ($p < 0.05$) as well as senior endoscopists ($p < 0.05$) across all five performance metrics: average accuracy, specificity, precision, recall, and F1 score, with statistically significant differences. Additionally, significant disparities are observed between DLKGS and the entire cohort of endoscopists in terms of accuracy, specificity, precision, recall, and F1 score ($p < 0.05$).

To give more details, recognition ability for each manifestation was compared between the DLKGS and endoscopists separately. The results of statistical significance analysis are shown in Table 7. One-sided t-tests show that DLKGS outperforms junior ($p < 0.05$), senior ($p < 0.05$) and all endoscopists combined across all metrics with statistically significant differences in recognizing DR, IM and N. In contrast, regarding to A and H recognition, DLKGS performs significantly better compared to junior endoscopists, but not compared to senior endoscopists.

4. Discussion

In this study, we developed a DL approach tailored for identifying the KCG from endoscopic images. We collected and annotated a dataset consisting of 29013 gastric endoscopic images from 2087 patients. Utilizing deep learning techniques, we trained DLKGS from these images specifically for KCG. The test results demonstrated that the model achieved an average accuracy,

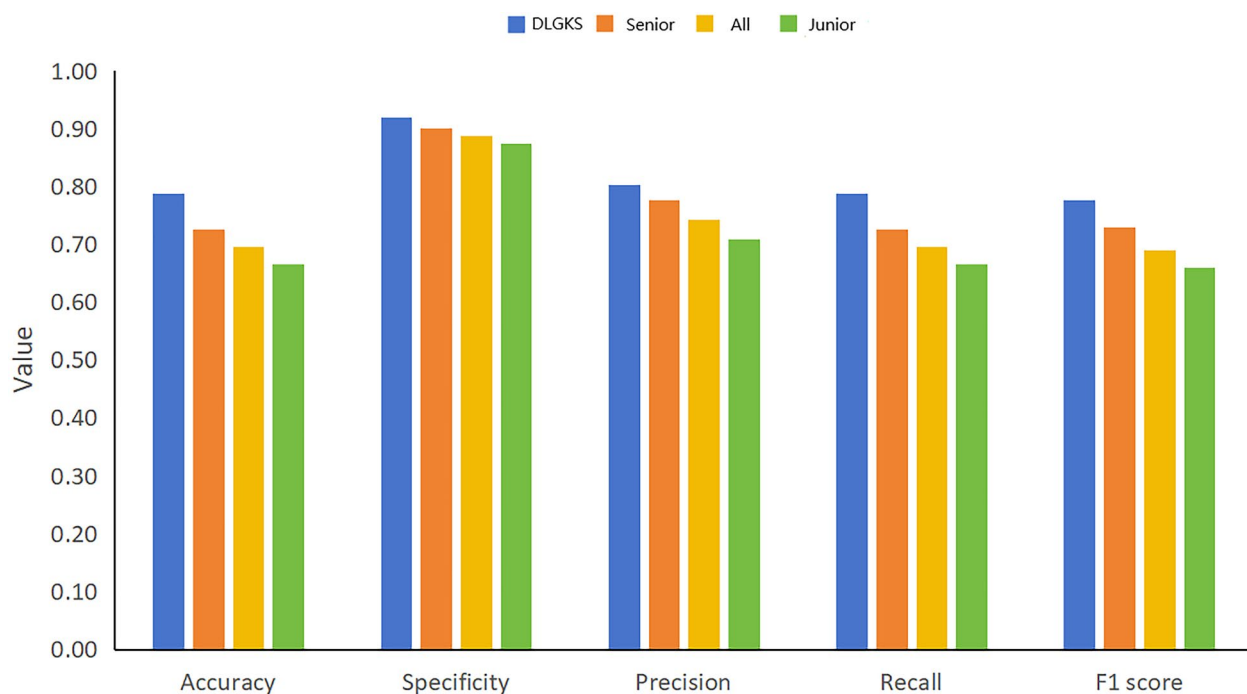


Figure 5. Comparison of gastroscopic manifestation recognition ability between DLKGS and endoscopists.

Table 6. Statistical significance analysis of the gastroscopic manifestation recognition ability between the DLGKS and endoscopists (* indicates $p < 0.05$).

Indicators	Comparison method	T-statistic	P value	95%CI	Mean difference
Accuracy (%)	DLGKS vs senior	5.198	*0.007*	[-3.24, 3.24]	6.07
	DLGKS vs junior	34.626	*<0.001*	[-0.97, 0.97]	12.10
	DLGKS vs all	7.847	*<0.001*	[-2.62, 2.62]	9.09
Specificity (%)	DLGKS vs senior	4.395	*0.012*	[-1.21, 1.21]	1.92
	DLGKS vs junior	34.617	*<0.001*	[-0.36, 0.36]	4.48
	DLGKS vs all	6.701	*<0.001*	[-1.08, 1.08]	3.20
Precision (%)	DLGKS vs senior	3.340	*0.029*	[-2.12, 2.12]	2.55
	DLGKS vs junior	22.853	*<0.001*	[-1.14, 1.14]	9.35
	DLGKS vs all	4.942	*0.001*	[-2.73, 2.73]	5.95
Recall (%)	DLGKS vs senior	5.198	*0.007*	[-3.24, 3.24]	6.07
	DLGKS vs junior	34.626	*<0.001*	[-0.97, 0.97]	12.10
	DLGKS vs all	7.847	*<0.001*	[-2.62, 2.62]	9.09
F1 Score	DLGKS vs senior	4.635	*0.010*	[-0.03, 0.03]	0.06
	DLGKS vs junior	47.357	*<0.001*	[-0.01, 0.01]	0.12
	DLGKS vs all	7.435	*<0.001*	[-0.03, 0.03]	0.09

Table 7. Statistical significance analysis of the ability in recognizing each manifestation individually between the DLGKS and endoscopists (* indicates $p < 0.05$).

Indicators	Comparison method	P value				
		A	DR	H	IM	N
Accuracy (%)	DLGKS vs senior	0.086	<0.001*	0.990	<0.001*	0.012*
	DLGKS vs junior	<0.001*	<0.001*	0.002*	<0.001*	0.001*
	DLGKS vs all	<0.001*	<0.001*	0.356	<0.001*	<0.001*
Specificity (%)	DLGKS vs senior	0.098	<0.001*	0.995	<0.001*	0.017*
	DLGKS vs junior	<0.001*	<0.001*	0.009*	<0.001*	<0.001*
	DLGKS vs all	<0.001*	<0.001*	0.536	<0.001*	<0.001*
Precision (%)	DLGKS vs senior	0.093	0.002*	0.988	0.003*	0.020*
	DLGKS vs junior	<0.001*	0.001*	0.005*	<0.001*	0.001*
	DLGKS vs all	<0.001*	<0.001*	0.308	<0.001*	<0.001*
Recall (%)	DLGKS vs senior	0.086	<0.001*	0.990	<0.001*	0.012*
	DLGKS vs junior	<0.001*	<0.001*	0.002*	<0.001*	0.001*
	DLGKS vs all	<0.001*	<0.001*	0.356	<0.001*	<0.001*
F1 Score	DLGKS vs senior	0.071	<0.001*	0.990	<0.001*	0.012*
	DLGKS vs junior	<0.001*	<0.001*	0.002*	<0.001*	0.001*
	DLGKS vs all	<0.001*	<0.001*	0.356	<0.001*	<0.001*

specificity, precision, recall, and F1 score of 78.70%, 91.92%, 80.23%, 78.70%, and 0.78, respectively. In comparison, the average performance of five senior endoscopists was 72.63%, 90.00%, 77.68%, 72.63%, and 0.73, while that of five junior endoscopists was 66.60%, 87.44%, 70.88%, 66.60%, and 0.66. Statistical analysis revealed that the model's performance significantly surpassed both the junior and senior endoscopists, indicating a notable accuracy advantage.

There are many other globally recognized systems for gastric cancer risk stratification, namely the Updated Sydney System (USS), Operative Link on Gastritis Assessment (OLGA), and Operative Link on Gastric Intestinal Metaplasia Assessment (OLGIM) [34–37]. However, these methods rely on multiple biopsies, leading to increased bleeding risk, pathology burden, and procedural time. These procedures also lack real-time prediction of gastric cancer risk, limiting their suitability for large-scale endoscopic screening. In contrast, KCG directly evaluates gastric cancer risk from endoscopic findings, demonstrating good histological consistency and wider clinical applicability [19, 38].

Moreover, KCG can offer real-time gastric cancer risk insights to endoscopists during endoscopic procedures, enhancing their alertness when examining high-risk patients.

Our study utilized a DL model, DLGKS, based on GAM-EfficientNet. In the field of digestive endoscopy, DLGKS has made significant breakthroughs in early cancer diagnosis. Tang et al. [39] investigated the feasibility of an AI model assisting in the diagnosis of early gastric cancer through narrow-band imaging endoscopy, achieving a test diagnostic accuracy of 93.2%, significantly higher than that of both senior (85.4%) and junior endoscopists (79.5%). Wu et al. [40] conducted a single-center, serial, randomized controlled trial, demonstrating that their developed AI system for identifying focal lesions and gastric cancer significantly decreased the missed diagnosis rate of gastric cancer, with a relative risk of 0.224. AI exhibits exceptional learning capabilities and advantages such as fatigue-free and standardized diagnosis. Integrating AI technology with KCG can promote the widespread use of KCG, improve gastric cancer risk screening

efficiency, enhance diagnostic consistency across regions and endoscopists, and facilitate further validation studies related to the KCG.

The KCG model developed in this study yielded satisfactory results in identifying and grading the five manifestations of Kyoto gastritis, demonstrating potential diagnostic value as a supplementary tool. However, there are limitations to consider: (1) a limited number of samples from a single source device, (2) uneven distribution of samples across multiple classifications, (3) inclusion of only five manifestations of the KCG, lacking recognition and differentiation abilities for other diseases, and (4) high image quality requirements. (5) Lack of validation with dataset of paired endoscopic images and histological reading. Future research will aim to conduct larger studies with more diverse samples, broaden the spectrum of diseases recognizable by the model, improve the recognition for low-quality images, and comprehensively validate the model with paired histological findings, with the goal to further enhance the accuracy and applicability of the model. In conclusion, this study has made preliminary achievements in AI-assisted KCG, demonstrating the significant potential of AI technology in improving the early diagnosis rate of gastric cancer. Despite the aforementioned limitations, we anticipate that with more data accumulation and ongoing model optimization, AI will increasingly contribute to gastric cancer risk screening and early diagnosis.

5. Conclusion

In this study, we developed a KGS approach, DLKGS, capable of identifying five manifestations of Kyoto gastritis, including atrophy, diffuse redness, enlarged folds, intestinal metaplasia, and nodular gastritis. DLKGS demonstrated superior performance compared to junior and senior endoscopists, and has potential to become a supplementary diagnostic tool for clinicians in gastroenterology practice.

Acknowledgments

We appreciate all coworkers from the Department of Digestive Endoscopy of Central Hospital of Dalian University for their assistance.

Authors contributions

Jing Zhang, Xiaoya Fan, and Zhong Wang designed the study; Ao Liu and Xilin Zhang drafted the manuscript, Jiaxin Zhong and Zilu Wang analyzed data; Zhenyang Ge and Zhong Wang revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Disclosure statement

Ao Liu, Xilin Zhang, Jiaxin Zhong, Zilu Wang, Zhenyang Ge, Zhong Wang, Xiaoya Fan, and Jing Zhang declare no conflict of interest to disclose.

Ethical approval and consent to participate

The present study was approved by the Medical Ethics Committee of Central Hospital of Dalian University of Technology (approval no. YN2022-047-06). All procedures were performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki and its later amendments. Signed informed consent form was obtained from each participant that permits the use of their data for research purposes.

Funding

This project was supported by the Fundamental Research Funds for the Central Universities (Grant Nos. DUT23YG236).

ORCID

Zhenyang Ge  <http://orcid.org/0000-0001-6535-8634>
Zhong Wang  <http://orcid.org/0009-0003-5602-1003>
Xiaoya Fan  <http://orcid.org/0000-0002-5002-6968>

Data availability statement

Due to the privacy of patients, all datasets generated and analyzed in the current study are not available unless a reasonable request to the correspondence author approved by the IRB of Central Hospital of Dalian University of Technology (J.Z., zj84402001@163.com).

References

- [1] Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA A Cancer J Clin.* 2021;71(3):209–249. doi:10.3322/caac.21660.
- [2] Cao W, Chen HD, Yu YW, et al. Changing profiles of cancer burden worldwide and in china: a secondary analysis of the global cancer statistics 2020. *Chin Med J (Engl).* 2021;134(7):783–791. doi:10.1097/CM9.0000000000001474.
- [3] Katai H, Ishikawa T, Akazawa K, et al. Five-year survival analysis of surgically resected gastric cancer cases in japan: a retrospective analysis of more than 100,000 patients from the nationwide registry of the Japanese gastric cancer association (2001–2007). *Gastric Cancer Off J Int Gastric Cancer Assoc Jpn Gastric Cancer Assoc.* 2018;21(1):144–154. doi:10.1007/s10120-017-0716-7.
- [4] Suzuki H, Ono H, Hirasawa T, et al. Long-term survival after endoscopic resection for gastric cancer: real-world evidence from a multicenter prospective cohort. *Clin Gastroenterol Hepatol Off Clin Pract J Am Gastroenterol Assoc.* 2023;21(2):307–318.e2. doi:10.1016/j.cgh.2022.07.029.

- [5] Smyth EC, Nilsson M, Grabsch HI, et al. Gastric cancer. *Lancet Lond Engl*. 2020;396(10251):635–648. doi:[10.1016/S0140-6736\(20\)31288-5](https://doi.org/10.1016/S0140-6736(20)31288-5).
- [6] Malfertheiner P, Megraud F, O'Morain CA, et al. Management of helicobacter pylori infection-the maas-tricht V/Florence consensus report. *Gut*. 2017;66(1):6–30. doi:[10.1136/gutjnl-2016-312288](https://doi.org/10.1136/gutjnl-2016-312288).
- [7] Banks M, Graham D, Jansen M, et al. British society of gastroenterology guidelines on the diagnosis and management of patients at risk of gastric adenocarcinoma. *Gut*. 2019;68(9):1545–1575. doi:[10.1136/gutjnl-2018-318126](https://doi.org/10.1136/gutjnl-2018-318126).
- [8] Pimentel-Nunes P, Libânio D, Marcos-Pinto R, et al. Management of epithelial precancerous conditions and lesions in the stomach (MAPS II): European society of gastrointestinal endoscopy (ESGE), european helicobacter and microbiota study group (EHMSG), european society of pathology (ESP), and sociedade portuguesa de endoscopia digestiva (SPED) guideline update 2019. *Endoscopy*. 2019;51(04):365–388. doi:[10.1055/a-0859-1883](https://doi.org/10.1055/a-0859-1883).
- [9] Zhang X, Li M, Chen S, et al. Endoscopic screening in asian countries is associated with reduced gastric cancer mortality: a meta-analysis and systematic review. *Gastroenterology*. 2018;155(2):347–354.e9. doi:[10.1053/j.gastro.2018.04.026](https://doi.org/10.1053/j.gastro.2018.04.026).
- [10] Yao K, Oishi T, Matsui T, et al. Novel magnified endoscopic findings of microvascular architecture in intramucosal gastric cancer. *Gastrointest Endosc*. 2002;56(2):279–284. doi:[10.1016/S0016-5107\(02\)70194-6](https://doi.org/10.1016/S0016-5107(02)70194-6).
- [11] Yao K, Iwashita A, Tanabe H, et al. Novel zoom endoscopy technique for diagnosis of small flat gastric cancer: a prospective, blind study. *Clin Gastroenterol Hepatol Off Clin Pract J Am Gastroenterol Assoc*. 2007;5(7):869–878. doi:[10.1016/j.cgh.2007.02.034](https://doi.org/10.1016/j.cgh.2007.02.034).
- [12] Doyama H, Nakanishi H, Yao K. Image-enhanced endoscopy and its corresponding histopathology in the stomach. *Gut Liver*. 2021;15(3):329–337. doi:[10.5009/gnl19392](https://doi.org/10.5009/gnl19392).
- [13] Dinis-Ribeiro M, da Costa-Pereira A, Lopes C, et al. Magnification chromoendoscopy for the diagnosis of gastric intestinal metaplasia and dysplasia. *Gastrointest Endosc*. 2003;57(4):498–504. doi:[10.1067/mge.2003.145](https://doi.org/10.1067/mge.2003.145).
- [14] Okubo M, Tahara T, Shibata T, et al. Usefulness of magnifying narrow-band imaging endoscopy in the Helicobacter pylori-related chronic gastritis. *Digestion*. 2011;83(3):161–166. doi:[10.1159/000321799](https://doi.org/10.1159/000321799).
- [15] Yamada M, Oda I, Taniguchi H, et al. Chronological trend in clinicopathological characteristics of gastric cancer. *Nihon Rinsho Jpn J Clin Med*. 2012;70(10):1681–1685.
- [16] Visaggi P, Barberio B, Gregori D, et al. Systematic review with meta-analysis: artificial intelligence in the diagnosis of oesophageal diseases. *Aliment Pharmacol Ther*. 2022;55(5):528–540. doi:[10.1111/apt.16778](https://doi.org/10.1111/apt.16778).
- [17] Wang F, Meng W, Wang B, et al. Helicobacter pylori-induced gastric inflammation and gastric cancer. *Cancer Lett*. 2014;345(2):196–202. doi:[10.1016/j.canlet.2013.08.016](https://doi.org/10.1016/j.canlet.2013.08.016).
- [18] Sugano K, Tack J, Kuipers EJ, et al. Kyoto global consensus report on helicobacter pylori gastritis. *Gut*. 2015;64(9):1353–1367. doi:[10.1136/gutjnl-2015-309252](https://doi.org/10.1136/gutjnl-2015-309252).
- [19] Toyoshima O, Nishizawa T, Yoshida S, et al. Consistency between the endoscopic kyoto classification and pathological updated sydney system for gastritis: a cross-sectional study. *J Gastro Hepatol*. 2022;37(2):291–300. doi:[10.1111/jgh.15693](https://doi.org/10.1111/jgh.15693).
- [20] Wang K, Zhao J, Jin H, et al. Establishment of a modified kyoto classification scoring model and its significance in the diagnosis of helicobacter pylori current infection. *Gastrointest Endosc*. 2023;97(4):684–693. doi:[10.1016/j.gie.2022.11.008](https://doi.org/10.1016/j.gie.2022.11.008).
- [21] Toyoshima O, Nishizawa T. Kyoto classification of gastritis: advances and future perspectives in endoscopic diagnosis of gastritis. *World J Gastroenterol*. 2022;28(43):6078–6089. doi:[10.3748/wjg.v28.i43.6078](https://doi.org/10.3748/wjg.v28.i43.6078).
- [22] Haruma K, Kato M, Inoue K, et al. Kyoto classification of gastritis. Tokyo: Nihon Medical Center; 2017.
- [23] Sugimoto M, Ban H, Ichikawa H, et al. Efficacy of the kyoto classification of gastritis in identifying patients at high risk for gastric cancer. *Intern Med*. 2017;56(6):579–586. doi:[10.2169/internalmedicine.56.7775](https://doi.org/10.2169/internalmedicine.56.7775).
- [24] Toyoshima O, Nishizawa T, Yoshida S, et al. Gastric cancer incidence based on endoscopic kyoto classification of gastritis. *World J Gastroenterol*. 2023;29(31):4763–4773. doi:[10.3748/wjg.v29.i31.4763](https://doi.org/10.3748/wjg.v29.i31.4763).
- [25] Yang H, Wu Y, Yang B, et al. Identification of upper GI diseases during screening gastroscopy using a deep convolutional neural network algorithm. *Gastrointest Endosc*. 2022;96(5):787–795.e6. doi:[10.1016/j.gie.2022.06.011](https://doi.org/10.1016/j.gie.2022.06.011).
- [26] Goto A, Kubota N, Nishikawa J, et al. Cooperation between artificial intelligence and endoscopists for diagnosing invasion depth of early gastric cancer. *Gastric Cancer Off J Int Gastric Cancer Assoc Jpn Gastric Cancer Assoc*. 2023;26(1):116–122. doi:[10.1007/s10120-022-01330-9](https://doi.org/10.1007/s10120-022-01330-9).
- [27] Wu L, Zhou W, Wan X, et al. A deep neural network improves endoscopic detection of early gastric cancer without blind spots. *Endoscopy*. 2019;51(06):522–531. doi:[10.1055/a-0855-3532](https://doi.org/10.1055/a-0855-3532).
- [28] Gazzah S, Amara NEB. New oversampling approaches based on polynomial fitting for imbalanced data sets. 2008 The Eighth IAPR International Workshop on Document Analysis Systems; 2008. IEEE. p. 677–684. doi:[10.1109/DAS.2008.74](https://doi.org/10.1109/DAS.2008.74).
- [29] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. IEEE. p. 770–778. doi:[10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90).
- [30] Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: synthetic minority over-sampling technique. *jair*. 2002;16:321–357. doi:[10.1613/jair.953](https://doi.org/10.1613/jair.953).
- [31] Tan M, Le Q. EfficientNet: rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning*; 2019. PMLR. p. 6105–6114. Accessed February 27, 2024. <https://proceedings.mlr.press/v97/tan19a.html>
- [32] Shi Y, Wei N, Wang K, et al. Deep learning-assisted diagnosis of chronic atrophic gastritis in endoscopy. *Front Oncol*. 2023;13:1122247. doi:[10.3389/fonc.2023.1122247](https://doi.org/10.3389/fonc.2023.1122247).
- [33] Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy; 2017. p. 618–626. doi:[10.1109/ICCV.2017.74](https://doi.org/10.1109/ICCV.2017.74).
- [34] Dixon MF, Genta RM, Yardley JH, et al. Classification and grading of gastritis. the updated sydney system. international workshop on the histopathology of gastritis, houston 1994. *Am J Surg Pathol*. 1996;20(10):1161–1181. doi:[10.1097/00000478-199610000-00001](https://doi.org/10.1097/00000478-199610000-00001).

- [35] Rugge M, Correa P, Di Mario F, et al. OLGA staging for gastritis: a tutorial. *Dig Liver Dis.* 2008;40(8):650–658. doi:[10.1016/j.dld.2008.02.030](https://doi.org/10.1016/j.dld.2008.02.030).
- [36] Rugge M, Genta RM. Staging and grading of chronic gastritis. *Hum Pathol.* 2005;36(3):228–233. doi:[10.1016/j.humpath.2004.12.008](https://doi.org/10.1016/j.humpath.2004.12.008).
- [37] Capelle LG, de Vries AC, Haringsma J, et al. The staging of gastritis with the OLGA system by using intestinal metaplasia as an accurate alternative for atrophic gastritis. *Gastrointest Endosc.* 2010;71(7):1150–1158. doi:[10.1016/j.gie.2009.12.029](https://doi.org/10.1016/j.gie.2009.12.029).
- [38] Na HK, Choi KD, Park YS, et al. Endoscopic scoring system for gastric atrophy and intestinal metaplasia: correlation with OLGA and OLGIM staging: a single-center prospective pilot study in Korea. *Scand J Gastroenterol.* 2022;57(9):1097–1104. doi:[10.1080/00365521.2022.2055974](https://doi.org/10.1080/00365521.2022.2055974).
- [39] Tang D, Ni M, Zheng C, et al. A deep learning-based model improves diagnosis of early gastric cancer under narrow band imaging endoscopy. *Surg Endosc.* 2022;36(10):7800–7810. doi:[10.1007/s00464-022-09319-2](https://doi.org/10.1007/s00464-022-09319-2).
- [40] Wu L, Shang R, Sharma P, et al. Effect of a deep learning-based system on the miss rate of gastric neoplasms during upper gastrointestinal endoscopy: a single-centre, tandem, randomised controlled trial. *Lancet Gastroenterol Hepatol.* 2021;6(9):700–708. doi:[10.1016/S2468-1253\(21\)00216-8](https://doi.org/10.1016/S2468-1253(21)00216-8).