International Journal of General Medicine

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Development and Validation of a Multi-Task Artificial Intelligence-Assisted System for Small Bowel Capsule Endoscopy

Jian Chen[1,2,*], Hongwei Wang[1,2,*], Zihao Zhang[3], Kaijian Xia[1,2], Fuli Gao[1], Xiaodan Xu[1], Ganhong Wang[4]

[1]Department of Gastroenterology, Changshu Hospital Affiliated to Soochow University, Suzhou, Jiangsu, 215500, People's Republic of China; [2]Center of Intelligent Medical Technology Research, Changshu Hospital Affiliated to Soochow University, Suzhou, 215500, People's Republic of China; [3]Department of Information Engineering, Shanghai Haoxiong Education Technology Co., Ltd, Shanghai, 200434, People's Republic of China; [4]Department of Gastroenterology, Changshu Hospital Affiliated to Nanjing University of Chinese Medicine, Suzhou, Jiangsu, 215500, People's Republic of China

*These authors have contributed equally to this work

Correspondence: Ganhong Wang; Xiaodan Xu, Email 651943259@qq.com; xxddocter@gmail.com

Objective: To develop a multi-task artificial intelligence-assisted system for small bowel capsule endoscopy (SBCE) based on various Transformer neural network architectures. The system integrates lesion recognition, cumulative time statistics, and progress bar marking functions to enhance the efficiency and accuracy of endoscopic image interpretation while effectively reducing missed diagnoses.

Methods: A dataset comprising 12 annotated categories of images captured by three different brands of capsule endoscopy devices was collected. Transfer learning and fine-tuning were conducted on five pre-trained Transformer models. Performance metrics, including accuracy, sensitivity, specificity, and recognition speed, were evaluated to select the best-performing model. The optimal model was converted from PyTorch to Open Neural Network Exchange (ONNX) format. Using OpenCV and MMCV tools, a multi-task SBCE-assisted reading system was developed.

Results: A total of 34,799 images were included in the study. The best-performing model, FocalNet, achieved a weighted average sensitivity of 85.69%, specificity of 98.58%, accuracy of 85.69%, and an AUC of 0.98 across all categories. Its diagnostic accuracy outperformed junior physicians (χ^2=17.26, p<0.05) and showed no statistical difference compared to senior physicians (χ^2=0.0716, p>0.05). The multi-task AI-assisted reading system, "FocalCE-Master", developed based on FocalNet, achieved a diagnostic speed of 592.40 frames per second, significantly faster than endoscopists. By integrating cumulative time bar charts with progress bar marking functionality, the system enables rapid localization and review of lesions, effectively streamlining the diagnostic workflow of SBCE.

Conclusion: The multi-task SBCE-assisted reading system developed using Transformer networks demonstrated rapid and accurate classification of various small bowel lesions. It holds significant potential in enhancing diagnostic efficiency and image review speed for endoscopists. However, the AI system has not yet been validated in prospective clinical trials, and further real-world studies are needed to confirm its clinical applicability.

Keywords: small bowel lesions, artificial intelligence, small bowel capsule endoscopy, transformer, transfer learning

Introduction

The small intestine, as a vital component of the digestive system, plays a key role in nutrient absorption and partial immune functions. However, due to its complex structure and considerable length, the clinical diagnosis of small intestine diseases has long been a challenge in the medical field. Traditional enteroscopy is time-consuming, highly invasive, and prone to lesion omission, limiting its widespread application. Since the advent of capsule endoscopy technology in 2000, its non-invasive nature, lack of anesthesia requirements, and high patient compliance have quickly

Received: 7 March 2025
Accepted: 7 May 2025
Published: 12 May 2025

made it the preferred method for diagnosing small intestinal diseases.[1,2] Although capsule endoscopy plays a vital role in diagnosing small intestinal lesions, a single examination typically generates between 40,000 to 60,000 images, significantly increasing the physician's reading workload and raising the risk of fatigue and missed diagnoses.[3,4]

In recent years, with the rapid advancement of artificial intelligence (AI) technologies, deep learning, particularly convolutional neural networks (CNNs), has provided highly efficient and accurate solutions for capsule endoscopy image interpretation. AI has demonstrated unique advantages in real-time object detection and lesion classification, significantly enhancing the efficiency and accuracy of lesion identification.[5–7] A systematic review and meta-analysis by Klang et al[8] further confirmed the high diagnostic performance of deep learning algorithms in wireless capsule endoscopy, highlighting their potential in ulcer, bleeding, and polyp detection tasks. For example, the study by Sabina Beg et al[9] developed a software called Omni Mode, which significantly reduced reading time by eliminating duplicate images. These systems, by shortening the time required for image review and reducing repetitive tasks, have not only substantially improved diagnostic efficiency but also effectively minimized the risks of missed and incorrect diagnoses, offering crucial support for the broader clinical adoption of capsule endoscopy. In recent years, the Transformer architecture, originally revolutionary in natural language processing, has been extended to the field of computer vision. Its self-attention mechanism enables efficient capture of long-range dependencies and a deeper understanding of complex data patterns. Compared to traditional CNNs, Transformers offer significant advantages in multitask learning, cross-domain applications, and generalization performance in complex scenarios.[10,11] Although Transformer models initially faced limitations in computational resource requirements, these challenges are being progressively addressed with the rapid improvement of GPU performance, the widespread adoption of hardware optimization technologies, and continuous advancements in model architectures.

Building on the strengths of Transformer models, this study developed an AI system for the automated recognition of multiple types of small bowel lesions. The system not only achieves accurate identification of various small bowel lesion categories but also innovatively integrates cumulative lesion duration statistics and progress bar category marking functions. This design visually presents the temporal distribution and occurrence timing of different lesions, providing comprehensive SBCE interpretation support for endoscopists. By combining these functionalities, the system is expected to further enhance the efficiency of assisted endoscopic image interpretation, reduce the workload and risk of missed diagnoses for physicians, and promote the broader application of AI technology in capsule endoscopy diagnostics, offering more intelligent solutions for clinical practice.

Methods

Study Design And datasets

This study utilized a total of four datasets, comprising 34,799 images that covered 12 categories of SBCE images (including one category of normal mucosal images). Examples of these images are shown in Figure 1. The images were captured using three different brands of capsule endoscopy devices: PillCam SB3 (Medtronic, USA), EndoCapsule (Olympus, Japan), and OMOM (Jinshan Science and Technology, China). The specific datasets included: Dataset #1 (SEE-AI open dataset), Dataset #2 (Kvasir-Capsule open dataset), Dataset #3 (Changshu Hospital of Traditional Chinese Medicine), and Dataset #4 (Changshu First People's Hospital). Dataset #1 was sourced from the SEE-AI open database, provided by Kyushu University Hospital, with images extracted from anonymized videos of patients undergoing small bowel CE examinations using Medtronic's PillCam SB3 (https://www.kaggle.com/datasets/capsuleyolo/kyucapsule?resource=download).[12] Dataset #2 was the Kvasir-Capsule open database, collected by Vestre Viken Hospital in Norway using the EndoCapsule system (https://osf.io/dv2ag/).[13] Datasets #3 and #4 were obtained by endoscopists from Changshu Hospital of Traditional Chinese Medicine and Changshu First People's Hospital, respectively, using OMOM capsule endoscopy devices from Jinshan Science and Technology. Datasets #1, #2, and #3 were stratified by lesion category and then split into training (n=27,036 images) and validation sets (n=6,750 images) in an 8:2 ratio to ensure balanced representation of each lesion type during model training. Dataset #4 was used exclusively as the test set, comprising 1,013 images and eight video clips.

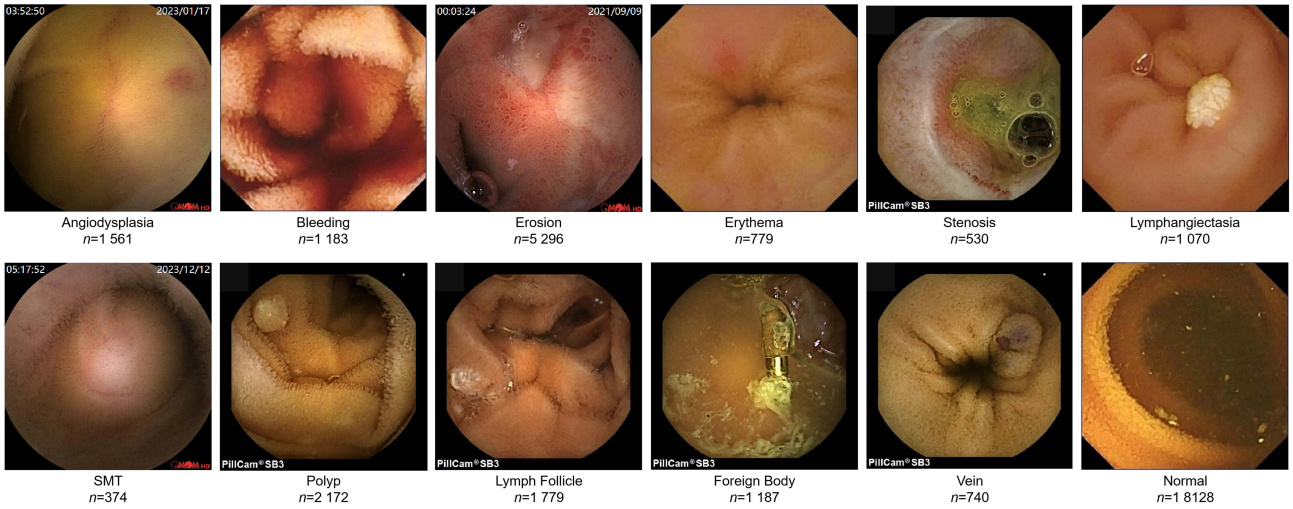

Figure 1 Examples of Images in the Dataset.

The image annotation process in this study was divided into three stages, with each stage managed by a dedicated team of endoscopists. In the Stage I (Figure 2A), an endoscopic physician selected video segments of interest and converted them into single-frame images. In the Stage II (Figure 2B), two additional teams of endoscopists screened these image frames, retaining clear images and those containing various lesions, while performing cross-checks to ensure accuracy. To evaluate annotation consistency, the Cohen's kappa coefficient was calculated between the two teams involved in Stage II, yielding a value of 0.82, which indicates a high level of agreement. In the Stage III (Figure 2C), a senior endoscopic physician reviewed the annotated results and made the final decisions. Figure 2 provides a detailed depiction of the image annotation workflow.

Image Preprocessing

During the training phase, image preprocessing strategies were tailored to meet the input size requirements of the selected pre-trained models. Input images were cropped and resized to target dimensions, such as 224×224 pixels or 256×256 pixels, to ensure compatibility with the model architecture. Specific preprocessing steps included random cropping and resizing to the target dimensions (RandomResizedCrop) and random horizontal flipping (RandomHorizontalFlip) to enhance data diversity. Additionally, images were converted into PyTorch tensor format

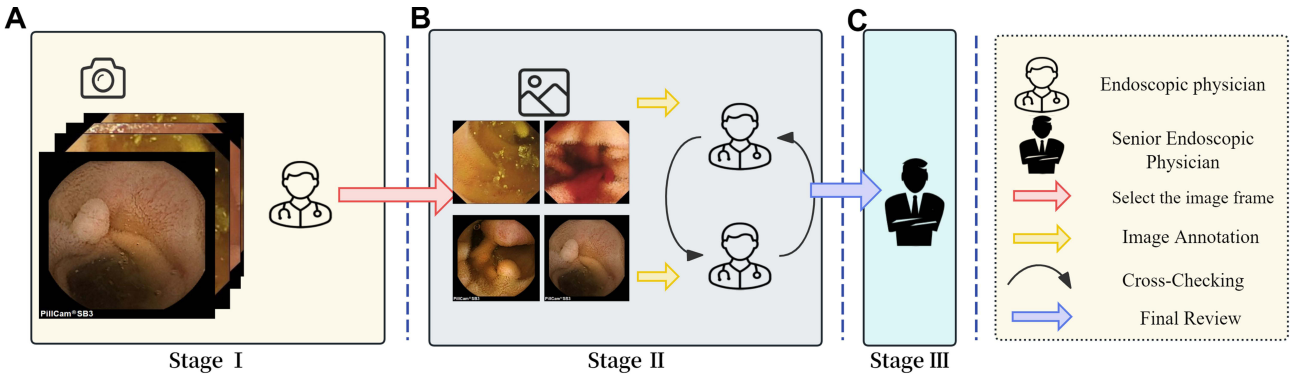

Figure 2 Workflow of Image Annotation. Workflow of Image Annotation. (**A**) Stage I: An endoscopic physician selected video segments and converted them into single-frame images. (**B**) Stage II: Two teams of endoscopists screened images, retained clear and lesion-containing frames, and cross-checked for accuracy. (**C**) Stage III: A senior endoscopic physician finalized the annotations.

https://doi.org/10.2147/IJGM.S522587

(ToTensor) and normalized using the ImageNet dataset's mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225] (Normalize) to reduce differences in image distribution.

During the testing phase, image preprocessing was similarly aligned with model requirements. Images were first resized to set the shorter edge to the target size (Resize), followed by cropping the target area at the center of the image (CenterCrop). The images were then converted to tensor format (ToTensor) and normalized using the same mean and standard deviation as in the training phase (Normalize). This ensured consistency in input images and stability in evaluation metrics.

Model Training Configuration

This study selected five models with different architectures pre-trained on the ImageNet dataset, including ViT (Vision Transformer),[14] CvT (Convolutional Vision Transformer),[15] Swin Transformer V2,[16] CaiT (Class-Attention in Image Transformers),[17] and FocalNet (Focal Transformer).[18] These models were trained using a transfer learning strategy. Pre-trained weight parameters were loaded, and all layers were fine-tuned on the constructed SBCE image dataset to meet the specific requirements of the SBCE recognition task. This approach leveraged the feature representation capabilities of ImageNet pre-trained weights while optimizing performance on the target dataset.

During the training phase, AdamW was employed as the optimizer with an initial learning rate of $1e^{-5}$. For most models (CvT, Swin Transformer V2, CaiT, and FocalNet), a StepLR scheduler was applied to dynamically adjust the learning rate every 7 epochs with a decay factor of γ=0.1, enhancing training stability. The ViT model, however, was trained with a constant learning rate without scheduling. The maximum number of training epochs was set to 60, with a batch size of 64 to ensure sufficient model learning. To improve training efficiency, an early stopping strategy was implemented. Training was automatically halted and the best-performing model saved when validation accuracy showed no improvement for 8 consecutive epochs. Detailed hyperparameter settings are provided in Table 1.

Multi-Task Capsule Endoscopy Reading System

This system utilizes various visualization tools to develop the trained Transformer model into a multi-task SBCE-assisted reading system. The main workflow includes model format conversion, model loading, video processing, frame prediction, result visualization, and video synthesis. First, the trained PyTorch model is converted into the Open Neural Network Exchange (ONNX) format to enable efficient cross-platform inference. The converted ONNX model is then loaded and set to evaluation mode. Simultaneously, a classification label mapping file is loaded to ensure that prediction results can be translated into their corresponding category names. Next, the MMCV tool is used to read input videos and process them frame by frame. Each frame undergoes image preprocessing, followed by model inference to generate category predictions and confidence scores, with the results annotated directly onto the image. Additionally, the system updates a bar chart in real-time to display the cumulative occurrence time for each category and shows a progress bar below the video. The progress bar uses color variations to represent the temporal distribution of classification results across the video. Once all frames are processed, the annotated and visualized frames are compiled into an output video. The system's architecture is illustrated in Figure 3.

Table 1 Model Hyperparameter Settings

Model	Learning Rate	Optimizer	Learning Rate Scheduler	Batch Size	Epochs	Patience
ViT	$1e^{-5}$	AdamW	StepLR (step=7)	64	60	8
CvT	$1e^{-5}$	AdamW	StepLR (step=7, γ=0.1)	64	60	8
Swin Transformer V2	$1e^{-5}$	AdamW	StepLR (step=7, γ=0.1)	64	60	8
CaiT	$1e^{-5}$	AdamW	StepLR (step=7, γ=0.1)	64	60	8
FocalNet	$1e^{-5}$	AdamW	StepLR (step=7, γ=0.1)	64	60	8

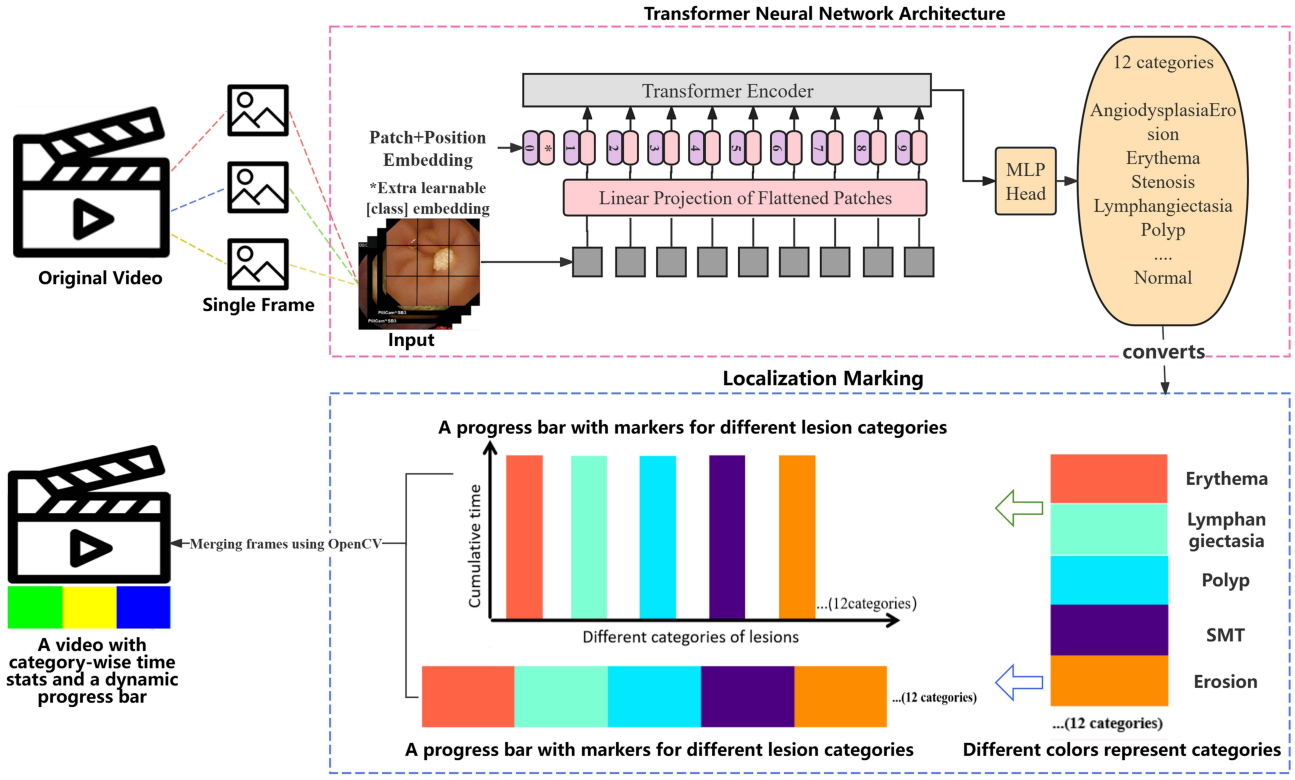

Figure 3 Architecture of the Transformer-Based Multi-Task SBCE-Assisted Reading System.

Model Interpretability Analysis

Although advanced computer vision technologies have achieved widespread application in medical imaging, high computational costs, data limitations, and the "black-box" nature of deep learning models remain significant barriers to their full adoption in the medical field. Similar to approaches in other studies, enhancing model transparency has become critical to fostering clinicians' trust in AI systems. To address this, explainable artificial intelligence (XAI) has been introduced, aiming to elucidate the internal workings and decision-making processes of deep learning models. To tackle this "black-box effect", this study conducted an in-depth interpretability analysis of the Transformer-based AI model, employing techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) and SHapley Additive exPlanations (SHAP).[19,20]

Grad-CAM generates class-specific localization maps by leveraging the gradient flow of the target concept (eg, a specific class) toward the final convolutional layer, highlighting important regions in the image. Mathematically, for a given classc, the gradient $\frac{\partial y^f}{\partial A^k}$ is computed with respect to the feature map A^k of the last convolutional layer. These gradients are globally averaged to obtain the weights α_k^c. The weighted sum of the feature maps is then passed through a ReLU activation to produce the output $L^c_{Grad-CAM} = ReLU\left(\sum_K \alpha_k^c A^k\right)$. SHAP (SHapley Additive exPlanations) values, derived from cooperative game theory, provide a unified measure of feature importance. In image classification, SHAP assigns importance to each pixel, indicating its contribution to the prediction. The SHAP value represents the difference in the model's expected output when a feature is included versus excluded, averaged across all possible combinations of features. This approach offers a comprehensive assessment of each feature's impact on the model's decision.

To better analyze the semantic-level classification capability of the model, intermediate layer outputs were extracted from the image classification model as semantic features. Different categories of small bowel lesions exhibit distinct

https://doi.org/10.2147/IJGM.S522587

semantic features. Forward hooks were set on the target layer to capture these features, which were then reduced to two-dimensional or three-dimensional space using t-SNE. The reduced features were visualized and analyzed using the Plotly library.[21]

Human-Machine Comparison

In the human-machine comparison experiment, two senior endoscopists with over five years of experience in SBCE image interpretation and two junior endoscopists with less than three years of experience were invited. They independently evaluated a test image set (n=1,013). Subsequently, the evaluation results of these endoscopists were compared with the image reading results of the model. The analysis focused on comparing the diagnostic accuracy and speed between the five Transformer models and endoscopists with varying levels of experience.

Experimental Platform and Evaluation Metrics

This study utilized a computer equipped with an NVIDIA GeForce RTX 4080 SUPER GPU (16GB VRAM), an Intel(R) Core(TM) i7-14700K CPU (3.4 GHz, 32GB RAM), and 1.9 TB SSD storage. The deep learning models were developed and trained using PyTorch (2.5.1), while image data processing was implemented with OpenCV (4.10.0.84). Data organization, analysis, and visualization were performed using Pandas (2.2.3), NumPy (2.0.2), Matplotlib (3.9.2), and Plotly (5.16.1), ensuring an efficient workflow and intuitive presentation of results. Additionally, the study integrated Weights & Biases (wandb, 0.18.7) to track and visualize metrics such as loss, accuracy, precision, sensitivity, and F1 score during training and testing. This ensured traceability of the experimental process and clear comparative insights into the results.

The AI model's performance was comprehensively evaluated using multiple metrics, including sensitivity, specificity, precision, accuracy, F1 score, average precision (AP), Matthews correlation coefficient (MCC), area under the receiver operating characteristic curve (AUC), Cohen's kappa coefficient (Cohen's Kappa), and weighted average. The calculation formulas are provided in Equations (1) to (8).

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \times 100\% \tag{1}$$

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\% \tag{2}$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100\% \tag{3}$$

$$\text{F1 score} = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity} \tag{4}$$

$$P_{weighted} = \sum_{i=1}^{k} w_i \cdot P_i \tag{5}$$

$$\text{average precision (AP)} : \text{AP} = \int_0^1 \text{p(r)dr} \tag{6}$$

$$\text{area under the receiver operating characteristic curve (AUC)} : \text{AUC} = 1/2 \sum_{\text{i=1}}^{\text{n-1}} (\text{FPR}_{\text{i+1}} - \text{FPR}_{\text{i}}) \times (\text{TPR}_{\text{i+1}} + \text{TPR}_{\text{i}}) \tag{7}$$

$$\text{Matthews Correlation Coefficient (MCC)} : \ \text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TN} + \text{FN})}} \tag{8}$$

In the study, TP represents the number of small intestinal lesions correctly identified as positive by the model, TN represents the number correctly identified as negative, FP denotes the number incorrectly identified as positive, and FN refers to the number incorrectly identified as negative. P_i is the performance metric value for the i category, and w_i represents the weight for the i category.

Table 2 Dataset Image Allocation and Distribution

Category	Training Set	Validation Set	Test Set	Total
Angiodysplasia	1159	289	113	1561
Bleeding	856	214	113	1183
Erosion	4172	1042	82	5296
Erythema	579	144	56	779
Foreign Body	868	216	103	1187
Lymph Follicle	1357	339	83	1779
Lymphangiectasia	812	202	56	1070
Normal Mucosa	14426	3603	99	18128
Polyp	1644	411	117	2172
SMT	268	66	40	374
Stenosis	382	96	52	530
Vein	513	128	99	740

Results

Baseline Characteristics

This study compiled four datasets, encompassing a total of 34,799 images captured by three different brands of capsule endoscopy (CE) devices. These images cover 12 categories of small intestinal lesions, including one category of normal mucosal images. Table 2 presents the distribution and allocation of images in the study, with 27,036 images in the training set, 6,750 in the validation set, and 1,013 in the test set.

Model Training

In this study, five different Transformer neural network architectures—ViT, CvT, Swin Transformer V2, CaiT, and FocalNet—were trained on the same dataset. During the training process, Weights & Biases (wandb) was utilized to monitor real-time changes in various metrics. The experimental results indicate that the classification loss for all five models decreased rapidly in the early stages of training as the number of training steps increased, eventually stabilizing, demonstrating effective convergence and optimization (Figure 4A). Simultaneously, the models' accuracy showed a significant improvement in the initial training phase and remained stable after reaching high levels (Figure 4B). Ultimately, all models achieved an accuracy exceeding 90%, with FocalNet performing the best, reaching a peak accuracy of 92.68%.

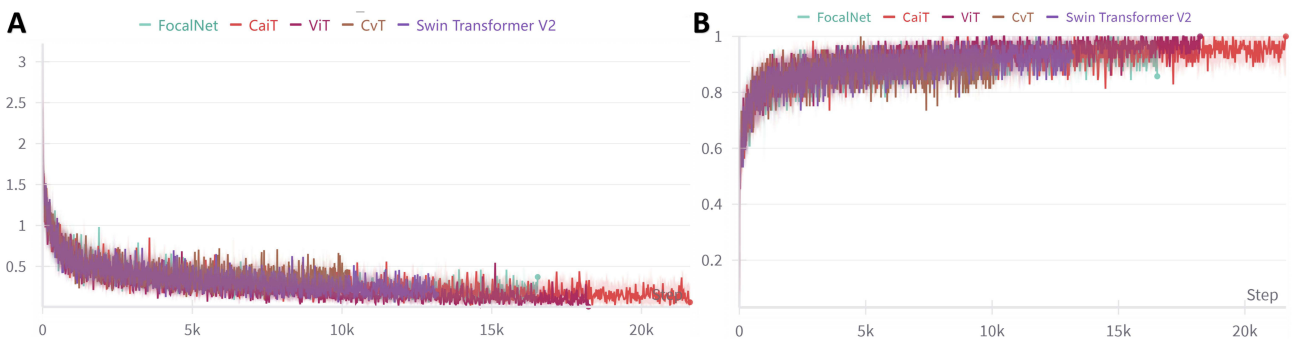

Figure 4 Trends in Metrics During Training Across Different Models. (**A**) Classification Loss: Shows the variation in classification loss with the number of training steps. (**B**) Accuracy: Illustrates the change in accuracy over training steps. Due to the varying complexity of model architectures, different batch sizes were used to maximize GPU performance and improve training efficiency, resulting in differences in the number of training steps among the models.

https://doi.org/10.2147/IJGM.S522587

Comparison of Diagnostic Performance Across Different Models

Table 3 compares the performance of five AI models trained via transfer learning—ViT, CvT, Swin Transformer V2, CaiT, and FocalNet—in the multi-class classification of small intestinal lesions using a validation set comprising 6,750 images. Among these models, FocalNet achieved the best overall performance, with an accuracy of 92.68%, sensitivity of 85.70%, F1 score of 86.13%, and frame rate of 593.11 frames per second, all of which were the highest values observed. Additionally, FocalNet demonstrated strong performance in precision, achieving 86.66% (ranking second). Consequently, FocalNet was selected as the optimal model.

Performance Evaluation of the Best Model on the Test Set

Table 4 provides a detailed evaluation of the best-performing model, FocalNet, on an external test set comprising 1,013 SBCE images. The table presents the performance metrics for 12 categories, including precision, sensitivity, specificity, F1 score, accuracy, average precision (AP), AUC, and Matthews Correlation Coefficient (MCC). Additionally, weighted averages are included as summary statistics.

Figure 5 illustrates two key evaluation curves for the predictive performance of the FocalNet model across different categories of small intestinal lesions on the test set: (A) Precision-Recall (PR) Curve and (B) ROC Curve. In Figure 5A, the PR curves for 11 of the 12 categories are close to the top-right corner, except for the "angiodysplasia" category, indicating superior predictive performance for these categories. In Figure 5B, the ROC curves demonstrate that, apart from the "polyp" category, where the curve does not fully approach the top-left corner, the curves for all other categories are tightly aligned with the top-left corner, further confirming the model's excellent classification performance for these categories.

Table 3 Performance Comparison of Different Models on the Validation Set

Model	Training Epochs	Accuracy (%)	Precision (%)	Sensitivity (%)	F1 Score (%)	Frame Rate (Frames/Second)
Swin Transformer V2	23	92.62	87.28*	84.86	85.96	506.54
CvT	15	90.67	83.64	79.26	81.06	552.25
CaiT	50	91.9	86.02	83.35	84.54	443.88
FocalNet	30	92.68*	86.66	85.70*	86.13*	593.11*
ViT	35	92.21	86.67	84.39	85.41	252.17

Notes: *Indicates the best performance. An early stopping strategy was applied, where training automatically halted, and the best model was saved if the validation set accuracy showed no improvement for 8 consecutive epochs. Consequently, the number of training epochs may vary.

Table 4 Performance Evaluation Results of the FocalNet Model on the Test Set

Category	Precision (%)	Sensitivity (%)	f1-score (%)	Specificity (%)	Accuracy (%)	AP (%)	AUC (95% CI)	MCC
Angiodysplasia	91.89	90.27	91.07	99.00	90.27	96.83	0.99(0.96, 1.00)	0.90
Bleeding	96.12	87.61	91.67	99.56	87.61	98.65	1.00(0.97, 1.00)	0.91
Erosion	64.81	85.37	73.68	95.92	85.37	82.78	0.98(0.93, 1.00)	0.72
Erythema	80.00	71.43	75.47	98.96	71.43	85.04	0.98(0.93, 1.00)	0.74
Foreign Body	99.04	100.00	99.52	99.89	100.00	100.00	1.00(0.97, 1.00)	0.99
Lymph Follicle	86.9	87.95	87.43	98.82	87.95	93.42	0.99(0.94, 1.00)	0.86
Lymphangiectasia	96.15	89.29	92.59	99.79	89.29	97.05	1.00(0.95, 1.00)	0.92
Normal Mucosa	62.82	98.99	76.86	93.65	98.99	96.18	0.99(0.94, 1.00)	0.76
Polyp	90.91	59.83	72.16	99.22	59.83	74.19	0.89(0.86, 0.92)	0.71
SMT	100.00	67.50	80.60	100.00	67.50	90.63	0.97(0.90, 1.00)	0.82
Stenosis	94.00	90.38	92.16	99.69	90.38	96.64	1.00(0.95, 1.00)	0.92
Vein	97.80	89.90	93.68	99.78	89.90	96.17	1.00(0.95, 1.00)	0.93
Weighted Average	88.12	85.69	85.84	98.58	85.69	92.30	0.98(0.95, 1.00)	0.85

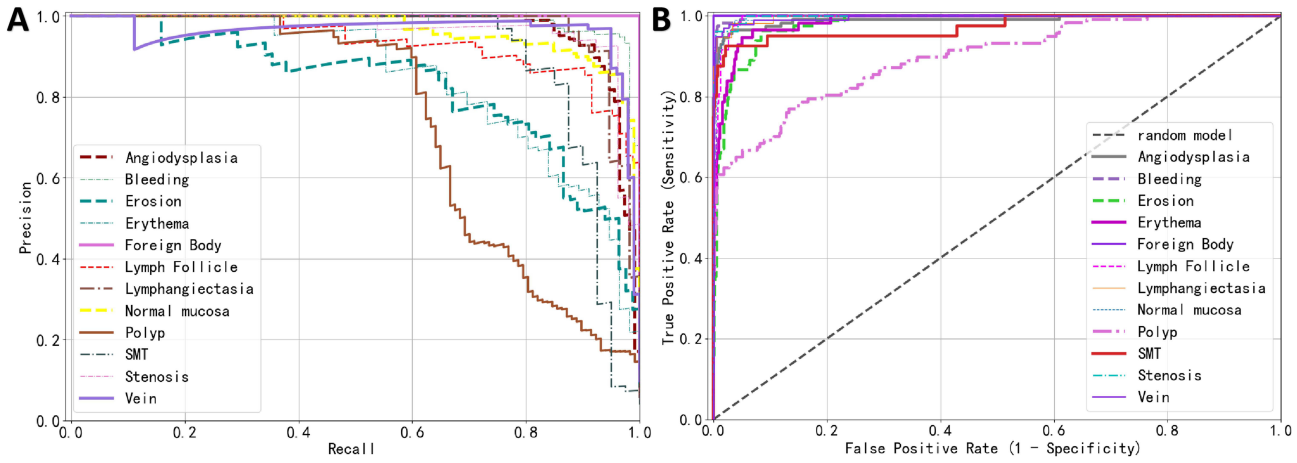

Figure 5 Predictive Performance of the AI Model on the External Test Set. (**A**) Precision-Recall (PR) Curve, (**B**) Receiver Operating Characteristic (ROC) Curve.

The following description provides an analysis of Figure 5, which highlights two critical evaluation curves for the FocalNet model's performance on the test set across various small intestinal lesion categories. (A) Precision-Recall (PR) Curve: The PR curves for 11 out of the 12 categories are positioned near the top-right corner of the graph, showcasing the model's superior predictive ability for these categories, with the exception of the "angiodysplasia" category. (B) ROC Curve: The ROC curves indicate exceptional classification performance for most categories, as they are tightly clustered near the top-left corner, except for the "polyp" category, which does not fully align with this trend. These results collectively demonstrate the model's robustness and high diagnostic accuracy across the majority of lesion types.

The effectiveness of the model's classification was analyzed using a confusion matrix to validate its accuracy and robustness across different categories, with detailed results presented in Figure 6A. The study demonstrates that the AI model performs exceptionally well in most cases. However, as shown in Figures 6B and C, some misclassifications still occur. These errors may be attributed to overlapping features between image categories, capsule movement, bubble interference, and image blurriness.

To investigate the causes of misclassification in SBCE lesion categorization by the AI model, t-SNE technology was employed to map high-dimensional data onto a two-dimensional plane, visually illustrating the separation between different categories (Figure 7). This approach aids in identifying easily distinguishable and overlapping categories, thereby uncovering the reasons for the model's misclassifications. Additionally, the study utilized t-SNE to construct an interactive 3D semantic feature map, accessible via an HTML file (https://pan.baidu.com/s/1bedUfcLCSuLhdpDBIPVGow?pwd=xxdd, password: xxdd). Users can intuitively explore each image and its position in the semantic feature space through mouse clicks, drags, and zooming. For instance, certain lesion categories, such as erosions and strictures, exhibit semantic feature overlap, which explains some of the model's misclassifications. This analysis provides valuable insights for model optimization and data augmentation strategies.

Model Visualization and Interpretation

Figure 8 illustrates the visualization of the decision-making process of the AI model using Grad-CAM technology. Figure 8A displays the original images; Figure 8B presents the pixel activation heatmaps generated based on the FocalNet model, highlighting the key regions influencing the model's decisions; Figure 8C overlays the activation heatmaps onto the original images, using warm colors (such as red and yellow) to indicate the critical lesion areas identified by the model. Each column includes examples from different lesion types.

Figure 9 utilizes SHAP technology to reveal the internal mechanisms of the model's predictive logic. In the two subplots, the model's predictions correspond to two true classifications: lymphangiectasia and varices. The pixel colors and their intensity indicate contributions to the model's predictions: red signifies positive contributions, blue indicates negative contributions, and deeper colors represent stronger impacts. In Figure 9A, compared to SMT, foreign body, and

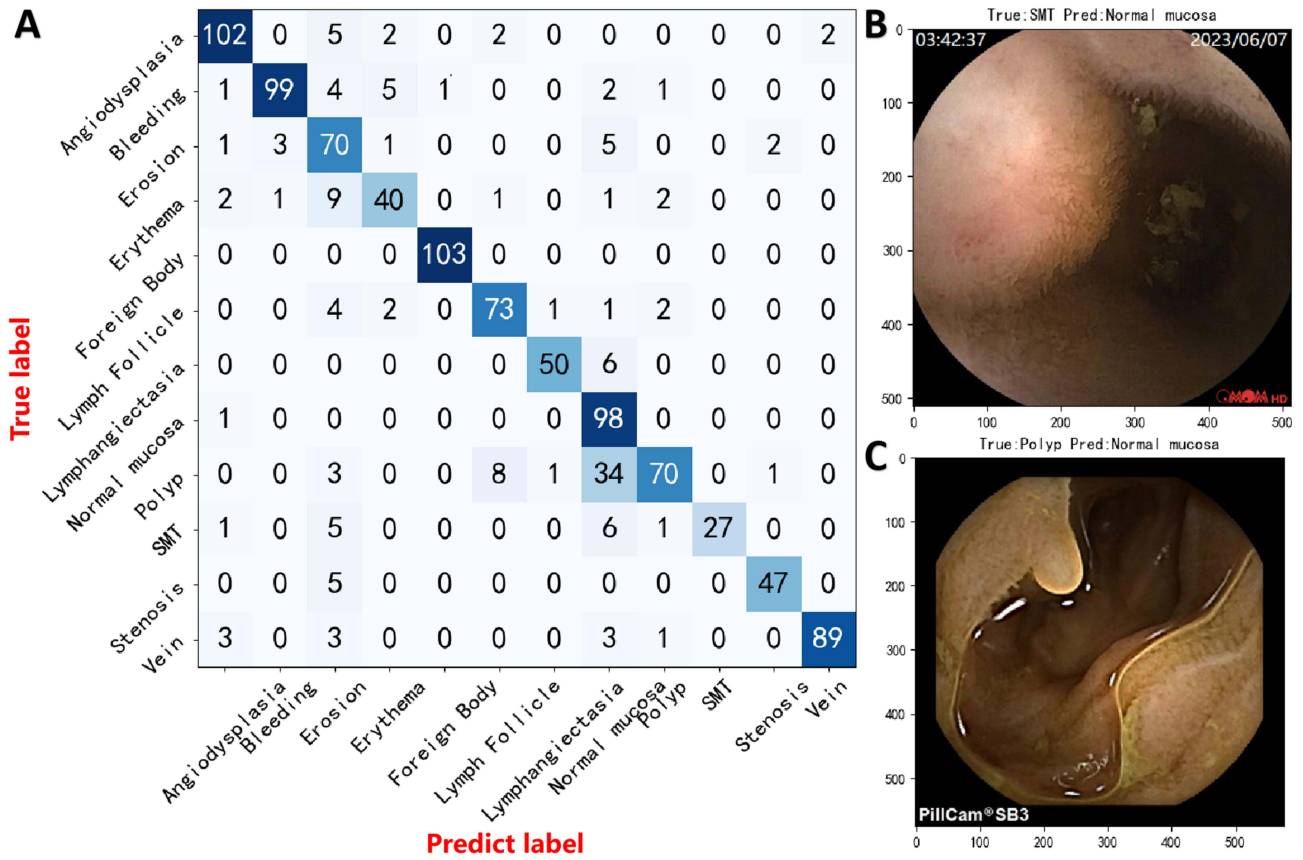

Figure 6 Model Performance on the Test Set. (**A**) Confusion Matrix: Demonstrates the classification accuracy of the model. (**B**) Misclassified Image Example: The model incorrectly classifies an image with the true label of submucosal tumor as normal mucosa. (**C**) Image Example: The true label is polyp, but it is mistakenly classified as normal mucosa.

erythema, the red regions for lymphangiectasia are more prominent, enabling the model to accurately classify it as lymphangiectasia. Similarly, the features highlighted in Figure 9B allow the model to correctly identify the case as vein.

Human-Machine Comparison Experiment

In this study, using a test set of 1,013 SBCE images, the diagnostic performance of five AI models was compared with that of endoscopists of varying experience levels, focusing on diagnostic accuracy and speed (measured in seconds). Among all models, the FocalNet model demonstrated the best performance, achieving a diagnostic accuracy of 85.69%, which was significantly higher than that of junior endoscopists (78.38%) and comparable to senior endoscopists (85.91%), as shown in Figure 10. χ^2-test results indicated that the diagnostic accuracy of the FocalNet model was significantly superior to that of junior endoscopists ($\chi^2 = 17.26$, $p < 0.05$) but showed no significant difference compared to senior endoscopists ($\chi^2 = 0.0716$, $p > 0.05$). These findings suggest that the FocalNet model outperforms junior endoscopists in diagnostic accuracy and performs on par with more experienced senior endoscopists. In terms of diagnostic speed, the FocalNet model required only 1.71 seconds to analyze 1,013 SBCE images (equivalent to 592.40 frames per second), which was markedly faster than both junior and senior endoscopists (Figure 10). The diagnostic speed of the AI model was approximately 357.43 times that of junior endoscopists and 335.20 times that of senior endoscopists.

Multi-Task AI-Assisted Image Analysis System

Our study developed a multi-task AI-assisted system for capsule endoscopy interpretation based on the optimal model FocalNet, named "FocalCE-Master." Figure 11A demonstrates the operation interface of the FocalCE-Master system,

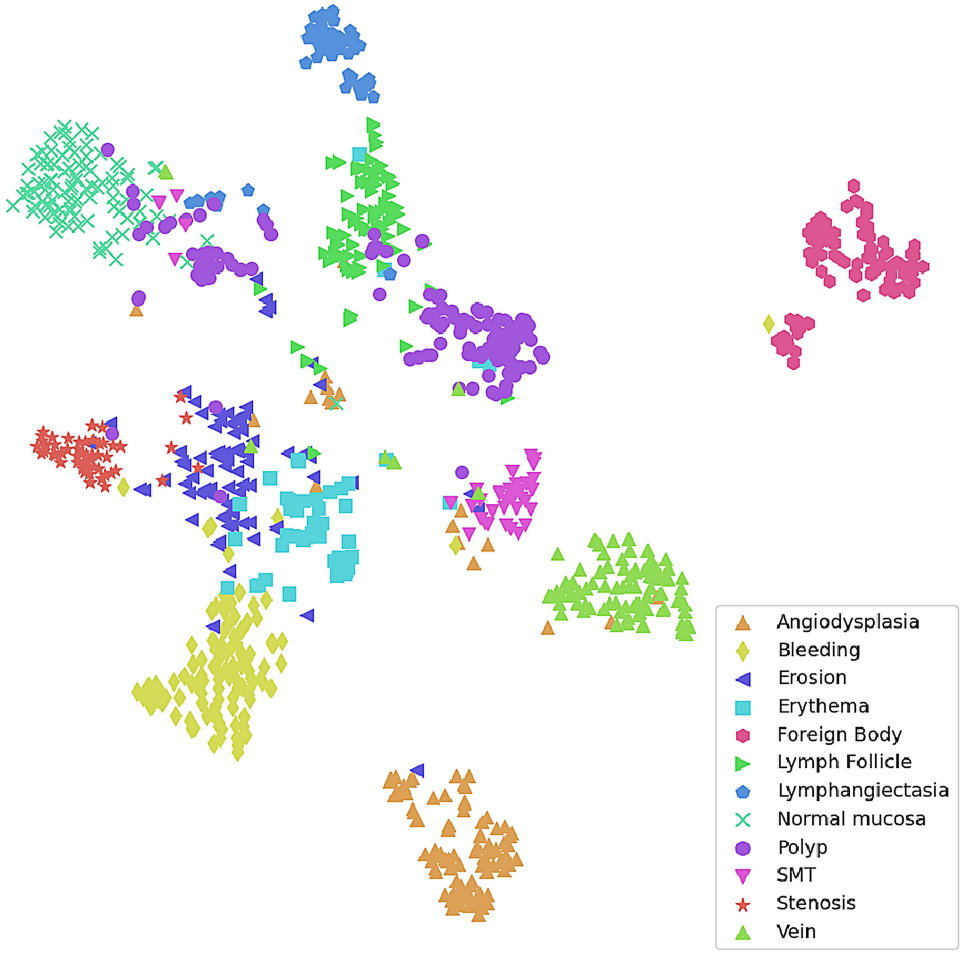

Figure 7 Two-Dimensional Semantic Feature Map of SBCE Images from the Test Set. Different colors and shapes represent various categories of small intestinal lesions, with similar features tending to cluster together on the map. If the boundaries between certain clusters are not well-defined, it indicates overlapping classifications in these regions, thereby increasing the difficulty of accurate categorization.

which consists of three sections: in the top left, the real-time view of the capsule endoscopy video is displayed, with the predicted category and its confidence score shown in red text at the upper left corner of the screen; in the top right, a bar chart is used to update the cumulative duration (in seconds) of each lesion type in the video in real time. The bar chart differentiates lesion types by color, with the x-axis representing lesion categories and the y-axis indicating cumulative time, and highlights the lesion type of the current frame; at the bottom, a color-coded progress bar illustrates the overall distribution of lesions throughout the video, with different colors corresponding to various lesion categories, aiding doctors in quickly locating areas of interest. This system integrates real-time prediction results, cumulative duration visualization, and progress bar lesion distribution to intuitively provide multi-dimensional information support for doctors in video analysis.

Case 1 (Figure 11B): In a video with a total duration of 4 minutes and 33 seconds, the capsule endoscopy captured multiple erosion lesions in the small intestine. FocalCE-Master accurately identified these lesions and labeled them in real time with the word “Erosion” in red text in the upper-left section of the screen. Users can view the cumulative duration of erosion lesions in the bar chart at the top right, while the specific time points of their appearance are clearly marked on the progress bar at the bottom. By dragging the progress bar to the marked positions, endoscopists can quickly review the frames showing erosion lesions, allowing for a comprehensive understanding of the lesions’ location and morphology, thereby improving diagnostic efficiency and accuracy. Case 2 (Figure 11C): This case demonstrates a small bowel stricture lesion. Through the real-time analysis and assistance provided by the FocalCE-Master system, endoscopists are similarly equipped with fast and accurate diagnostic support.

https://doi.org/10.2147/IJGM.S522587

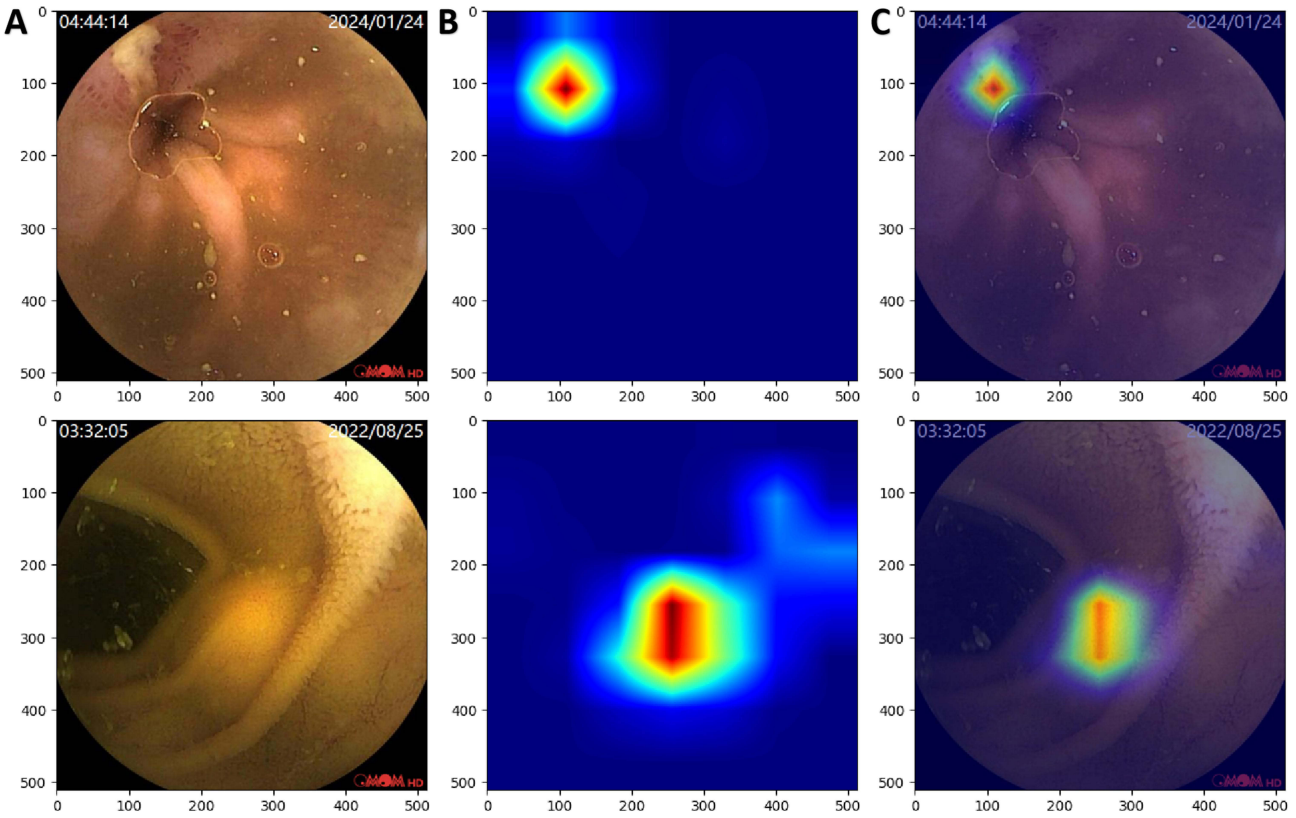

Figure 8 Grad-CAM Visualization of the AI Model's Decision-Making Process. Column (**A**) displays the original endoscopic images, Column (**B**) shows the pixel activation heatmaps generated using Grad-CAM, and Column (**C**) features the overlay of the original images and activation heatmaps.

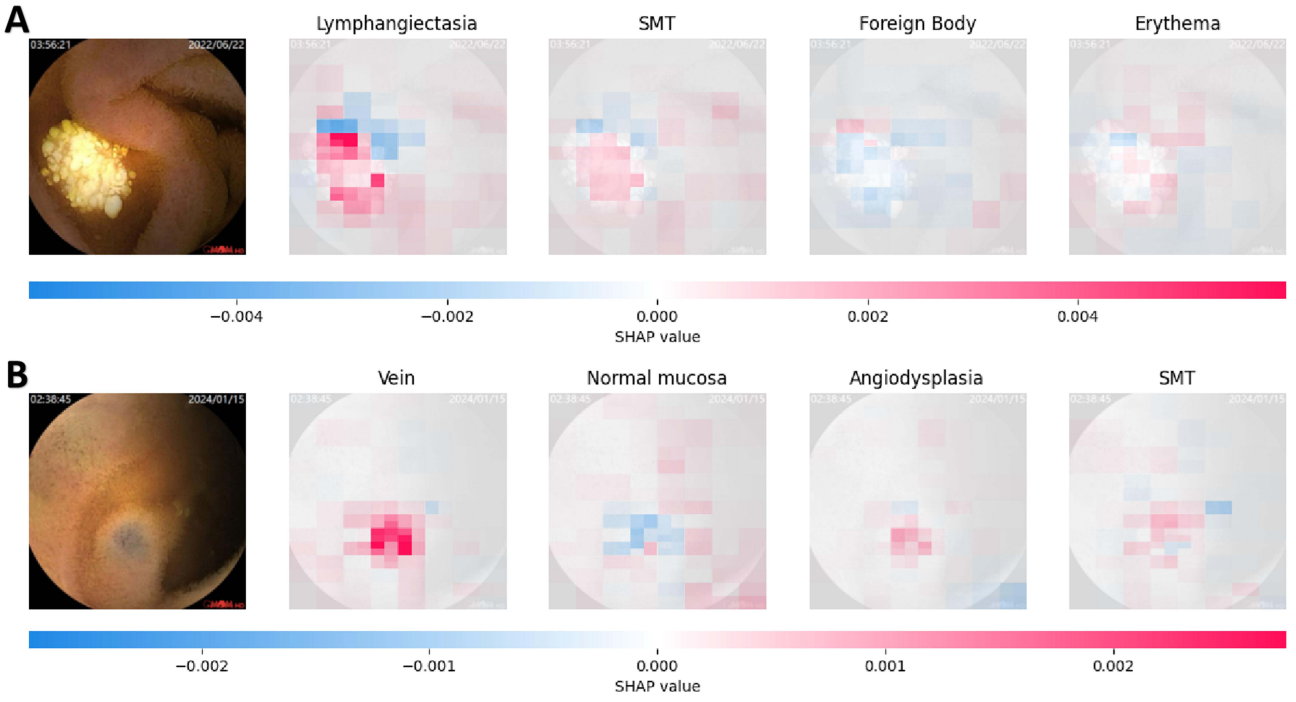

Figure 9 An Interpretability Analysis Based on SHAP Technology. (**A**) Capsule Endoscopy Image Correctly Predicting "Lymphangiectasia": The model accurately identifies the "Lymphangiectasia" label using SHAP values. (**B**) Capsule Endoscopy Image Correctly Predicting "Vein": The model precisely determines the "Vein" label through SHAP values. SHAP Value Significance: SHAP values represent the magnitude of each pixel's contribution to the classification outcome. Higher values indicate greater importance of the pixel in the prediction. Positive values (depicted in red) signify a supportive impact, while negative values (illustrated in blue) denote an inhibitory effect.

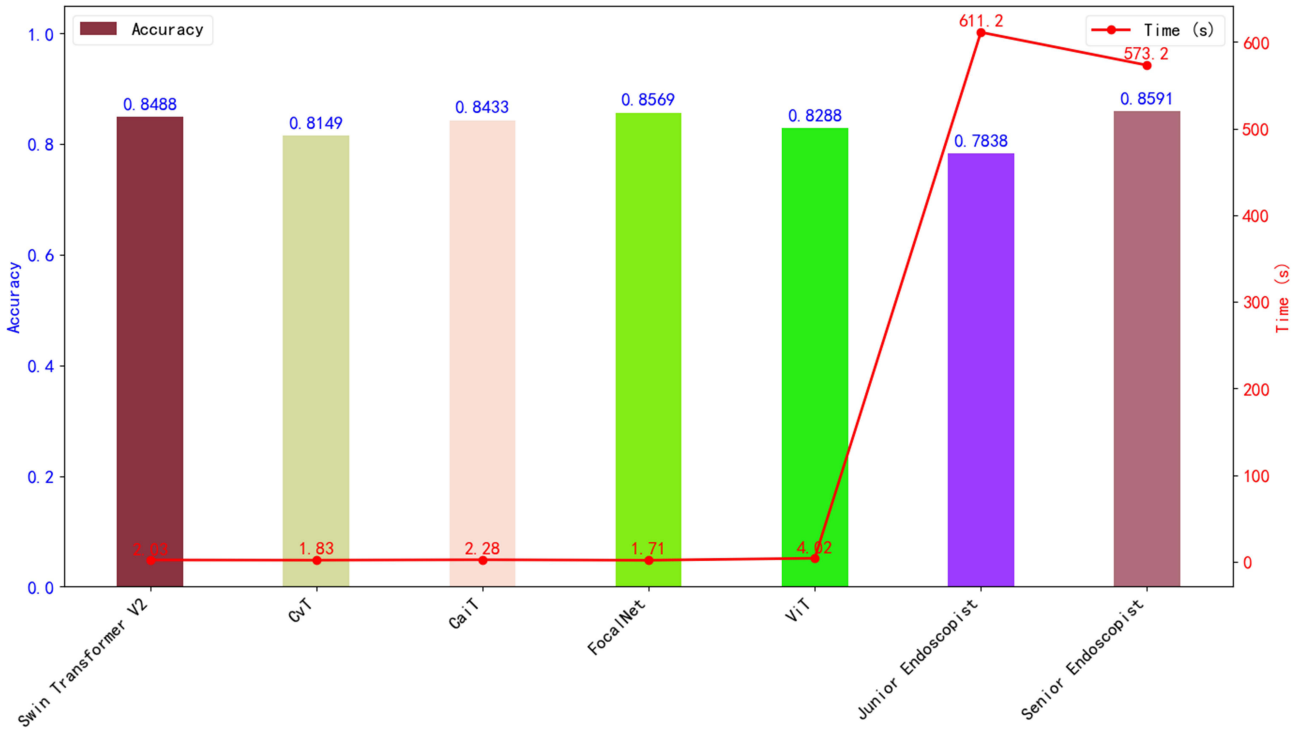

Figure 10 Comparison of Diagnostic Performance Between AI Models and Endoscopists. This figure illustrates the diagnostic accuracy and speed of AI models compared to endoscopists with varying levels of experience. The bar chart represents the comparison of diagnostic accuracy, while the line chart depicts the comparison of diagnostic time (in seconds). The left vertical axis indicates accuracy rates, and the right vertical axis represents diagnostic time.

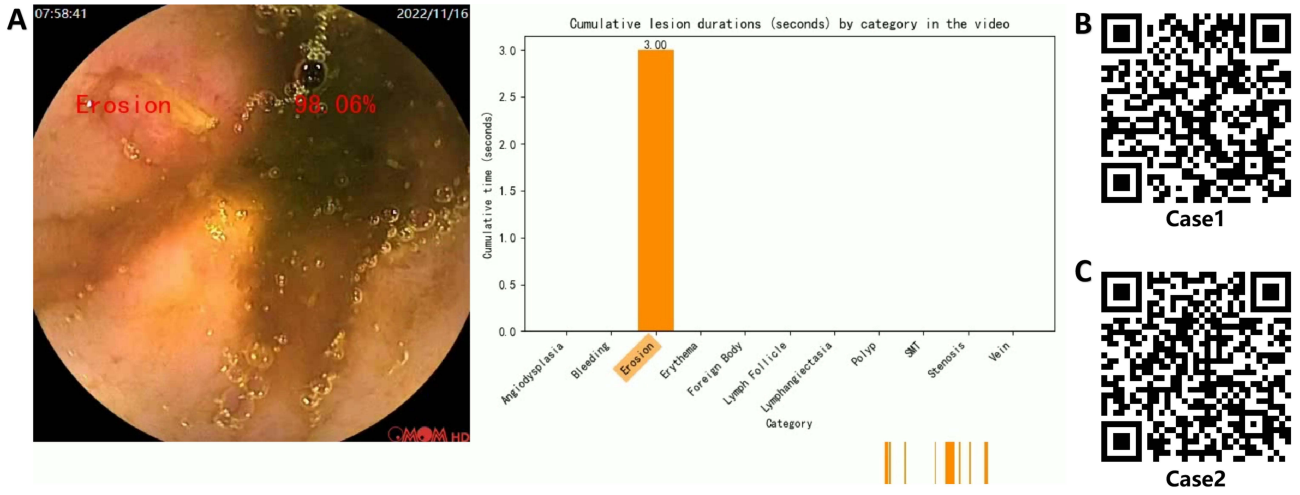

Figure 11 Multi-task AI-assisted interpretation system. (**A**) Operational interface of the multi-task AI-assisted interpretation system developed in this study; (**B** and **C**) showcase two case examples of image interpretation using this system.

Discussion

This study utilized five different Transformer model architectures for training, validation, and testing on a specialized dataset of SBCE images and videos captured by three different brands of capsule devices from four distinct medical centers. Among the models compared, FocalNet demonstrated the best performance in terms of accuracy (92.68%), sensitivity (85.70%), F1 score (86.13%), and recognition speed (593.11 frames per second). Its recognition accuracy was significantly higher than that of junior endoscopists ($\chi^2 = 17.26$, $p < 0.05$) and showed no significant difference compared

https://doi.org/10.2147/IJGM.S522587

to senior endoscopists (85.91%, $\chi^2 = 0.0716$, $p > 0.05$). In terms of diagnostic speed, FocalNet was approximately 357.43 times faster than junior endoscopists and 335.20 times faster than senior endoscopists. Furthermore, the best-performing model was developed into a multi-task AI-assisted system named FocalCE-Master, equipped with lesion recognition, cumulative time statistics, and progress bar marking functions, further enhancing its clinical utility.

The advent of capsule endoscopy has made direct visualization of the small intestinal mucosa possible, marking a major breakthrough in endoscopic technology over the past three decades.[22] Compared with traditional small bowel endoscopy, capsule endoscopy offers simpler operation and less invasiveness, whereas small bowel endoscopy requires both oral and anal approaches to complete the observation of the entire small intestine.[23] Compared to multi-slice CT enterography, capsule endoscopy also has the advantage of being radiation-free. Despite its convenience, capsule endoscopy faces challenges in practical application due to the labor-intensive process of manual interpretation: physicians must examine a vast amount of image data frame by frame, which is time-consuming and exhausting, increasing the risk of missed diagnoses. This is particularly concerning as lesion areas typically account for only 5% to 10% of the images, sometimes even less. To address these issues, computer-assisted intelligent detection and recognition of small bowel lesions is emerging as a new trend to support clinical diagnosis. The proposed FocalCE-Master system demonstrates a significant advantage in diagnostic speed, requiring only 1.71 seconds to analyze 1,013 capsule endoscopy images (equivalent to 592.40 frames per second), far exceeding the interpretation speed of junior and senior endoscopists. Considering that a single capsule endoscopy examination typically generates 40,000 to 60,000 images, FocalCE-Master, with its processing speed of 592.40 frames per second, can complete the analysis of 60,000 images in approximately 1.7 minutes. This highlights its outstanding capability in rapid diagnosis of large-scale SBCE images, significantly improving diagnostic efficiency and practicality. In addition, this study employed Grad-CAM and SHAP visualization tools to intuitively illustrate the decision-making process of the AI model. These visualizations not only help clinicians better understand the model's recognition mechanisms but also serve as educational aids to enhance the ability of junior endoscopists to identify small bowel lesions.

Previous studies have primarily focused on developing AI-assisted models for small bowel lesion detection. For instance, Saito et al[24] developed an AI model based on deep convolutional neural networks to detect elevated small bowel lesions. Similarly, Yokote et al[12] utilized the YOLOv5 model to construct an AI system capable of recognizing 12 types of annotated lesion images, while also making the related dataset publicly available. The FocalCE-Master system proposed in this study is a multi-task auxiliary system that not only performs real-time inference and prediction for each frame in capsule endoscopy videos but also converts different lesion categories into corresponding color-coded cumulative time bar charts and progress bars. In practical applications, capsule endoscopists can use the cumulative time bar chart to intuitively understand the lesion categories involved throughout the examination and their cumulative durations. For lesions of interest, clinicians can drag the slider on the progress bar, marked with the lesion's color, to quickly locate and review video segments corresponding to the lesion's time frame. This enables a comprehensive understanding of the lesion's distribution and characteristics. This design significantly improves diagnostic efficiency and accuracy while providing robust auxiliary support for the analysis of complex cases.

This study developed an AI-assisted interpretation system based on a dataset from four medical centers, capable of recognizing 12 categories of small bowel lesion images (including one category of normal mucosa). The dataset encompassed images from multiple capsule systems (PillCam, EndoCapsule, OMOM) and centers, enhancing model robustness and generalizability. While this study focused on diagnostic accuracy and interpretability, future work should further assess performance across diverse populations and device types to address potential biases. Compared to previous systems that could only identify single lesion types, such as bleeding[25] or elevated lesions,[26] the FocalCE-Master system demonstrates significant advantages in practicality. By supporting multi-category lesion recognition, this system is better suited to the diverse demands of complex clinical scenarios, providing comprehensive diagnostic support for small bowel lesions. In terms of weighted average performance across all categories, the FocalCE-Master system achieved a sensitivity of 85.69%, accuracy of 85.69%, precision of 88.12%, and an F1 score of 85.84%, indicating high reliability in overall diagnostic performance. However, performance varied across different lesion categories. Notably, sensitivity

was relatively low for polyps and submucosal tumors, at 59.83% and 67.50%, respectively, falling short of the results reported by Jian C et al[5] (74.36% and 72.50%, respectively). The observed variations in lesion-specific sensitivity may be attributed to semantic feature overlaps among certain lesion categories, as illustrated by the t-SNE analysis. Such overlaps likely led to misclassification between lesions with similar visual characteristics, particularly between polyps and lymphangiectasia.

Conclusions

This study developed the multi-task AI-assisted system FocalCE-Master based on Transformer models to address the challenges of low efficiency and high missed diagnosis risk in manual capsule endoscopy interpretation. Trained and tested on a multi-brand capsule endoscopy dataset from four medical centers, FocalCE-Master demonstrated excellent performance with an accuracy of 92.68%, sensitivity of 85.70%, F1 score of 86.13%, and recognition speed of 593.11 frames per second. Its diagnostic speed was significantly faster than both junior and senior endoscopists, while its diagnostic accuracy surpassed that of junior endoscopists. The system supports the recognition of 12 categories of small bowel lesions and enables rapid lesion localization and review through cumulative time bar charts and progress bar functionalities. These features significantly enhance the system's ease of use and clinical practicality, making it a valuable tool for real-world applications. However, due to the limited number of images related to Crohn's disease and parasitic infections, the current dataset remains insufficient to support the development of a robust and reliable AI model. In the future, expanding the dataset through multicenter data sharing may enhance the model's generalizability and robustness. Moreover, to effectively incorporate FocalCE-Master into routine clinical workflows, several deployment challenges must be addressed, including seamless integration with existing clinical information systems, data interoperability, clinician acceptance, as well as subsequent external validation and regulatory compliance processes (eg, CE/FDA certification). These issues represent critical steps that must be carefully managed before large-scale clinical implementation can be realized.

Ethics Approval and Consent to Participate

This study has obtained approval from the Ethics Committee of Changshu Hospital Affiliated to Soochow University (the IRB approval number L2024054). This study was performed in accordance with the Declaration of Helsinki, and written informed consent was obtained from all participants.

Author Contributions

Jian Chen and Hongwei Wang contributed equally to this work and share first authorship. All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This study received financial support from the Changshu City Science and Technology Plan Project (CS202452, CSWS202316); the Suzhou Health Information and Medical Big Data Society Project (SZMIA2402); the Health Informatics Key Support Discipline Funding of Suzhou City (SZFCXK202147); and Suzhou City's 23rd Science and Technology Development Plan Project (SLT2023006). No funding body had any role in the design of the study and collection, analysis, interpretation of data, or in writing the manuscript.

Disclosure

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

1. Hosoe N, Takabayashi K, Ogata H, Kanai T. Capsule endoscopy for small-intestinal disorders: current status. *Digestive Endoscopy*. 2019;31(5):498–507. doi:10.1111/den.13346
2. Lazaridis L-D, Tziatzios G, Toth E. Implementation of European Society of Gastrointestinal Endoscopy (ESGE) recommendations for small-bowel capsule endoscopy into clinical practice: results of an official ESGE survey. *Endoscopy*. 2021;53(9):970–980. doi:10.1055/a-1541-2938
3. Beg S, Card T, Sidhu R, Wronska E, Ragunath K. The impact of reader fatigue on the accuracy of capsule endoscopy interpretation. *Digestive Liver Dis*. 2021;53(8):1028–1033. doi:10.1016/j.dld.2021.04.024
4. Park J, Hwang Y, Yoon J, et al. Recent development of computer vision technology to improve capsule endoscopy. *Clin Endosc*. 2019;52(4):328–333. doi:10.5946/ce.2018.172
5. Chen J, Xia K, Zhang Z, Ding Y, Wang G, Xu X. Establishing an AI model and application for automated capsule endoscopy recognition based on convolutional neural networks (with video). *BMC Gastroenterol*. 2024;24(1):394. doi:10.1186/s12876-024-03482-7
6. Ding Z, Shi H, Zhang H. Gastroenterologist-level identification of small-bowel diseases and normal variants by capsule endoscopy using a deep-learning model. *Gastroenterology*. 2019;157(4):1044–1054. doi:10.1053/j.gastro.2019.06.025
7. Ding Z, Shi H, Zhang H. Artificial intelligence-based diagnosis of abnormalities in small-bowel capsule endoscopy. *Endoscopy*. 2023;55(1):44–51. doi:10.1055/a-1881-4209
8. Soffer S, Klang E, Shimon O, et al. Deep learning for wireless capsule endoscopy: a systematic review and meta-analysis. *Gastrointest Endosc*. 2020;92(4):831–839. doi:10.1016/j.gie.2020.04.039
9. Beg S, Wronska E, Araujo I. Use of rapid reading software to reduce capsule endoscopy reading times while maintaining accuracy. *Gastrointest Endosc*. 2020;91(6):1322–1327. doi:10.1016/j.gie.2020.01.026
10. Ali AM, Benjdira B, Koubaa A, El-Shafai W, Khan Z, Boulila W. Vision Transformers in Image Restoration: a Survey. *Sensors*. 2023;23(5).
11. Aymen F, Monir H, Pester A. Large vision models: how transformer-based models excelled over traditional deep learning architectures in video processing. *Qatar Med J*. 2024;2024(4):50–54. doi:10.5339/qmj.2024.50
12. Yokote A, Umeno J, Kawasaki K, et al. Small bowel capsule endoscopy examination and open access database with artificial intelligence: the SEE-artificial intelligence project. *DEN Open*. 2024;4(1). doi:10.1002/deo2.258.
13. Smedsrud PH, Thambawita V, Hicks SA, et al. Kvasir-Capsule, a video capsule endoscopy dataset. *Sci Data*. 2021;8(1):142. doi:10.1038/s41597-021-00920-z
14. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale *Int Conference on Learning Representations(ICLR)* 2021.11929
15. Wu H, Xiao B, Codella N, et al. CvT: introducing Convolutions to Vision Transformers. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada. 2021;22–31.
16. Liu Z, Hu H, Lin Y, et al. Swin transformer v2: scaling up capacity and resolution. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* New Orleans, LA, USA. 2022;11999–12009.
17. Touvron H, Cord M, Sablayrolles A, Synnaeve G, Jégou H. Going deeper with image transformers. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, QC, Canada 2021;32–42.
18. Yang J, Li C, Dai X, Gao J. Focal Modulation Networks *Adv Neural Information Processing Systems (Neurips)* 2022.
19. Zhang Y, Hong D, McClement D, Oladosu O, Pridham G, Slaney G. Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging. *J Neurosci Meth*. 2021;353:109098. doi:10.1016/j.jneumeth.2021.109098
20. Kikutsuji T, Mori Y, Okazaki K, Mori T, Kim K, Matubayasi N. Explaining reaction coordinates of alanine dipeptide isomerization obtained from deep neural networks using explainable artificial intelligence (XAI). *J Chem Phys*. 2022;156(15):154108. doi:10.1063/5.0087310
21. Linderman GC, Steinerberger S. Clustering with t-SNE, provably. *Siam J Math Data Sci*. 2019;1(2):313–332. doi:10.1137/18M1216134
22. Wang A, Banerjee S, BA B, et al. Wireless capsule endoscopy. *Gastrointest Endosc*. 2013;78(6):805–815. doi:10.1016/j.gie.2013.06.026
23. Pennazio M, Rondonotti E, Despott EJ, et al. Small-bowel capsule endoscopy and device-assisted enteroscopy for diagnosis and treatment of small-bowel disorders: European Society of Gastrointestinal Endoscopy (ESGE) Guideline - Update 2022. *Endoscopy*. 2023;55(1):58–95. doi:10.1055/a-1973-3796
24. Saito H, Aoki T, Aoyama K, et al. Automatic detection and classification of protruding lesions in wireless capsule endoscopy images based on a deep convolutional neural network. *Gastrointest Endosc*. 2020;92(1):144–151. doi:10.1016/j.gie.2020.01.054
25. Musha A, Hasnat R, Mamun AA, Ping EP, Ghosh T. Computer-aided bleeding detection algorithms for capsule endoscopy: a systematic review. *Sensors*. 2023;23(16):7170. doi:10.3390/s23167170
26. Kim HJ, Gong EJ, Bang CS, Lee JJ, Suk KT, Baik GH. Computer-aided diagnosis of gastrointestinal protruded lesions using wireless capsule endoscopy: a systematic review and diagnostic test accuracy meta-analysis. *J Pers Med*. 2022;12(4).644