BMC Gastroenterology

RESEARCH

Open Access

AI support for colonoscopy quality control using CNN and transformer architectures

Jian Chen[1†], Ganhong Wang[2†], Jingjie Zhou[1], Zihao Zhang[3], Yu Ding[1], Kaijian Xia[4*] and Xiaodan Xu[1*]

Abstract

Background Construct deep learning models for colonoscopy quality control using different architectures and explore their decision-making mechanisms.

Methods A total of 4,189 colonoscopy images were collected from two medical centers, covering different levels of bowel cleanliness, the presence of polyps, and the cecum. Using these data, eight pre-trained models based on CNN and Transformer architectures underwent transfer learning and fine-tuning. The models' performance was evaluated using metrics such as AUC, Precision, and F1 score. Perceptual hash functions were employed to detect image changes, enabling real-time monitoring of colonoscopy withdrawal speed. Model interpretability was analyzed using techniques such as Grad-CAM and SHAP. Finally, the best-performing model was converted to ONNX format and deployed on device terminals.

Results The EfficientNetB2 model outperformed other architectures on the validation set, achieving an accuracy of 0.992. It surpassed models based on other CNN and Transformer architectures. The model's precision, recall, and F1 score were 0.991, 0.989, and 0.990, respectively. On the test set, the EfficientNetB2 model achieved an average AUC of 0.996, with a precision of 0.948 and a recall of 0.952. Interpretability analysis showed the specific image regions the model used for decision-making. The model was converted to ONNX format and deployed on device terminals, achieving an average inference speed of over 60 frames per second.

Conclusions The AI-assisted quality system, based on the EfficientNetB2 model, integrates four key quality control indicators for colonoscopy. This integration enables medical institutions to comprehensively manage and enhance these indicators using a single model, showcasing promising potential for clinical applications.

Keywords Deep learning, Colonoscopy quality control, Colonoscopy, Artificial intelligence

†Jian Chen and Ganhong Wang contributted equally to this work.

*Correspondence:
Kaijian Xia
kjxia@suda.edu.cn
Xiaodan Xu
xxddocter@gmail.com
[1]Department of Gastroenterology, Changshu Hospital Affiliated to Soochow University, Suzhou 215500, China
[2]Department of Gastroenterology, Changshu Traditional Chinese Medicine Hospital (New District Hospital), Suzhou 215500, China
[3]Shanghai Haoxiong Education Technology Co., Ltd, Shanghai 200434, China
[4]Department of Information Engineering, Changshu Hospital Affiliated to Soochow University, Suzhou 215500, China

Introduction

Colorectal cancer (CRC) is the second leading cause of cancer-related deaths worldwide [1]. This disease may begin with non-cancerous polyps in the colon, which, if not treated in time, can develop into cancer [2]. Typically, CRC progresses through the "adenoma-carcinoma" sequence, with polyps undergoing malignant transformation over several years, during which they can be detected and treated [3]. Therefore, early screening, detection, and timely removal of polyps are crucial in reducing the incidence of colorectal cancer.

With the continuous advancement of gastrointestinal endoscopy technology, colonoscopy has become the preferred method for early CRC screening and the diagnosis of colonic lesions [4]. Identifying and promptly removing adenomatous polyps can significantly reduce the risk of colorectal cancer [5]. However, the cleanliness of the bowel is crucial for accurate examination results, as poor bowel preparation can lead to missed lesions [6, 7]. Complete cecal intubation is also a key factor in the quality of colonoscopy [8, 9]. Slowing the withdrawal speed has been shown to significantly increase the adenoma detection rate (ADR)and reduce the risk of interval colorectal cancer [10, 11]. In recent years, the European Society of Gastrointestinal Endoscopy, the Digestive Endoscopy Society of the Chinese Medical Association, and the American Society for Gastrointestinal Endoscopy have issued quality control statements on colonoscopy screening [12–14], highlighting polyp detection rate, bowel preparation quality, withdrawal speed, and cecal intubation rate as critical indicators for colonoscopy quality control.

Deep learning, with its exceptional feature extraction and data processing capabilities, offers intelligent solutions for colonoscopy quality control, especially in the evaluation of polyps, bowel preparation, and the cecum [15, 16]. Convolutional Neural Networks (CNN) mainly handle data of fixed shapes, such as images, whereas Transformers based on self-attention have set new standards in natural language processing and have expanded into the computer vision domain [17]. This study employs these deep learning architectures with the aim to intelligently evaluate key quality indicators of colonoscopy, providing doctors with real-time feedback and supplying data for further training, thereby enhancing the diagnostic and therapeutic efficacy of colonoscopy.

Convolutional Neural Networks (CNNs) primarily handle fixed-shape data such as images, while Transformers, based on self-attention mechanisms, have not only set new standards in natural language processing but have also expanded into the field of computer vision [17]. Both approaches, with their exceptional feature extraction and data processing capabilities, have found extensive applications in gastrointestinal endoscopy, aiding endoscopists in enhancing the efficiency and accuracy of diagnoses. Karaman et al. [18] proposed a new method for optimizing activation functions and hyperparameters of the YOLOv5 algorithm for real-time detection of colorectal polyps. They developed an AI-based real-time monitoring system to oversee withdrawal speed during colonoscopy. Gong et al. [19] found that the proportion of over-speed frames (POF) during colonoscope withdrawal is negatively correlated with the adenoma detection rate (ADR), meaning that a lower POF is associated with a higher ADR. Additionally, deep learning technology can automatically evaluate bowel cleanliness before colonoscopy, standardizing and enhancing the accuracy of such assessments and reducing variability in human evaluations, thereby improving diagnostic accuracy and efficiency [20].

In this study, we developed an artificial intelligence-assisted system that integrates four key quality control indicators for colonoscopy to enable automated diagnosis. These indicators include real-time monitoring of withdrawal speed, improving polyp detection rates, automatic assessment of bowel preparation quality, and ensuring cecal intubation rate, all recommended as critical factors for colonoscopy quality by several international guidelines. This system allows medical institutions to comprehensively manage and enhance these indicators through a single model.

Methods

Study design and datasets

This study was based on two datasets: Dataset 1 (HyperKvasir) was used for model training and validation, while Dataset 2 (Changshu Hospital Affiliated to Soochow University) served as an external test set for the model. The collected colonoscopy images covered intestines of varying cleanliness (using the Boston Bowel Preparation Score, BBPS), polyps, and the cecum. HyperKvasir (Dataset 1) [21] is the largest gastrointestinal endoscopy dataset (https://datasets.simula.no/hyper-kvasir/). It contains over 110,079 images and 374 videos. These data were collected during real gastroscopy and colonoscopy examinations at Bærum Hospital in Norway, with some marked by experienced gastroenterological endoscopists. This dataset represents anatomical landmarks as well as pathological and normal findings. From this, we selected 1009 cecum images, 1028 polyp images, and 1794 images of intestines with different levels of cleanliness for analysis. In addition, we retrospectively collected 358 colonoscopy images from the Endoscopy Center of Changshu Hospital affiliated with Soochow University as an external test set (Dataset 2). Relevant image examples can be seen in Fig. 1. To enhance the model's generalization, the collected endoscopy images utilized various image-enhancing endoscopy techniques, such as Narrow Band Imaging

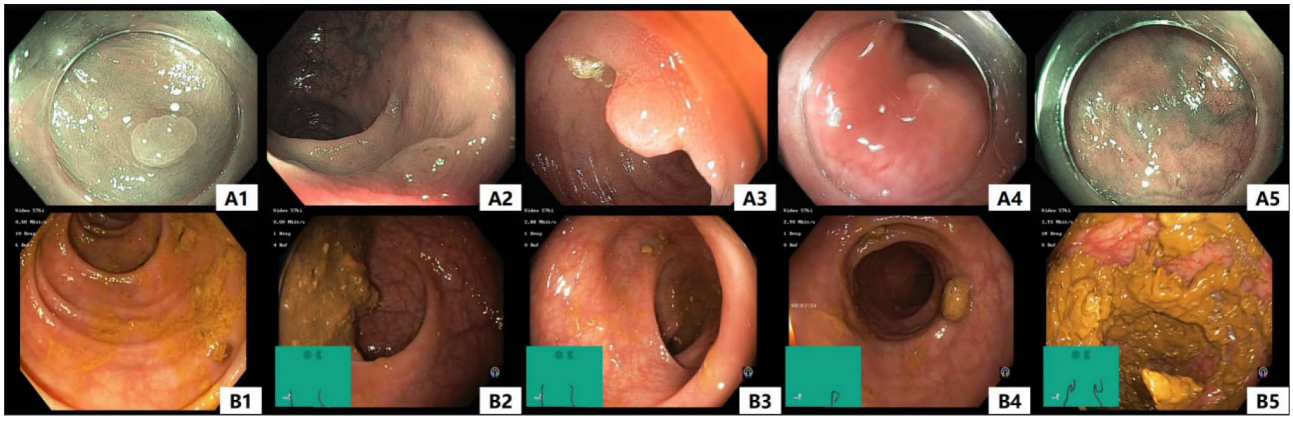

Fig. 1 Representative images from the dataset; (**A**) Polyp images A1-A6; (**B**) Intestinal images for BBPS scores B1-B6. BBPS: Boston Bowel Preparation

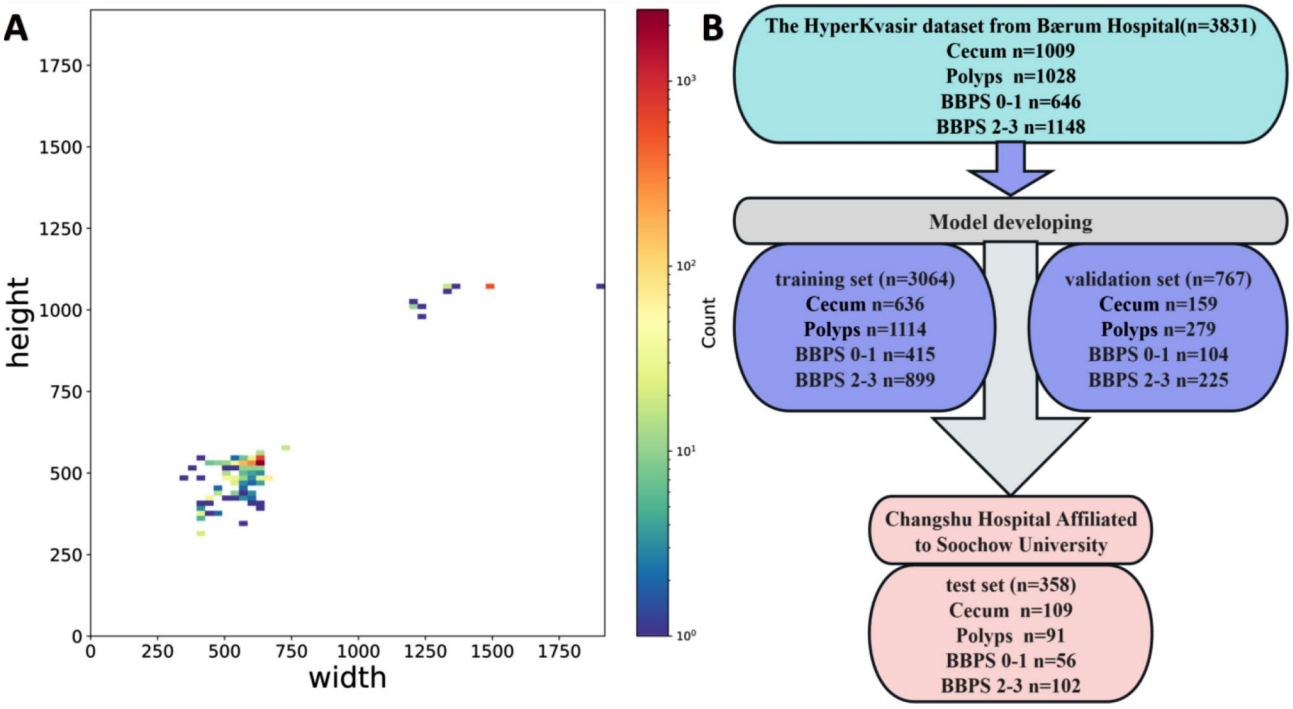

Fig. 2 Distribution of images in the datasets; (**A**) Distribution of image sizes. Red indicates a higher concentration of images of that size, while blue indicates fewer. (**B**) Distribution of image categories in the training set, validation set, and test set

(NBI), Blue Light Imaging (BLI), and Flexible Spectral Imaging Color Enhancement (FICE).

This study excluded patients with the following conditions: inflammatory bowel disease, active colitis, coagulation disorders, familial polyposis, emergency colonoscopy, and those with incomplete diagnostic and treatment information. Figure 2A displays the image size distribution of the two datasets. Notably, the dataset includes images of various sizes, among which those with dimensions of 622×529 and 633×532 together account for more than 50% of the total. The distribution of images across different categories in the training set, validation set, and test set can be seen in Fig. 2B.

Construction of the AI system

Image preprocessing

In our study, to ensure enhanced model generalization, we implemented a series of preprocessing and augmentation methods on the image data. The distribution of images across the training, validation, and test sets is detailed in Table 1. For the training set, we began by randomly resizing the images and cropping them to a 224×224 dimension. To diversify the data, we introduced random horizontal flips. Following this, we transitioned images from either PIL Image or numpy.ndarray format to PyTorch Tensor, normalizing their range to [0, 1]. In the final stages, we standardized the RGB channels of

Table 1 Image distribution in training, validation, and test sets

class	Train set	Validation Set	Test Set
bbps 0–1	415	104	56
bbps 2–3	899	225	102
cecum	636	159	109
polyps	1114	279	91

the images, employing means of [0.485, 0.456, 0.406] and standard deviations of [0.229, 0.224, 0.225]. For the test set, we adopted a varied approach, first adjusting the image's shorter edge to 256 pixels, and subsequently performing a centered 224×224 crop. The subsequent transformations and normalizations mirrored those applied to the training set, utilizing the same RGB channel standardization parameters. Each of these steps was executed using PyTorch's torchvision library.

Model training configuration

For image classification, we leveraged pre-trained models rooted in both Convolutional Neural Network (CNN) and Transformer deep learning architectures for transfer learning. Within the CNN framework, we opted for models including DenseNet-121, EfficientNetB2, ResNet50, and VGG19 (Visual Geometry Group Network). Meanwhile, within the Transformer structure, we employed the ViT (Vision Transformer), Swin (Shifted Window Transformer), DeiT(Data-efficient Image Transformers), and CvT (Convolutional Vision Transformer) models. These CNN models encompass convolutional layers, average pooling layers, and fully connected layers with ReLU activations. To tailor to our dataset, two dense layers with ReLU activations were appended to each pre-trained model, along with an output layer with Softmax activation for classification. The number of features in the output layer was set to four, aligning with our classification objectives. Models utilized cross-entropy as the loss function and underwent 30 epochs of training with the Adam optimizer. Concurrently, we enacted a learning rate schedule, halving the rate every five epochs. When processing input images, Transformer models initiated with random cropping, horizontal flips, and rotations up to 15 degrees. The models then segmented the images into fixed-size patches, adding positional encodings for each patch. These patches were addressed in the Transformer encoder to ascertain inter-patch relationships, and only the output from the first patch was used for the four-class classification. All procedures were conducted within the PyTorch framework. For a detailed view of the neural network architectures, refer to Fig. 3. The core operation of a CNN is convolution, defined as: $(\mathrm{f}*\mathrm{g})(\mathrm{t}) = \int_{-\infty}^{\infty} \mathrm{f}(\tau)\,\mathrm{g}(\mathrm{t}-\tau)\,\mathrm{d}\tau$,where f is the input image, g is the convolution kernel, and t represents the pixel coordinates.The self-attention mechanism is a key component of Transformers, represented by the formula:

$$\mathrm{Attention}\,(\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{softmax}\left(\frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{\mathrm{d_k}}}\right)\mathrm{V}$$

where $\mathrm{Q}\,\mathrm{K}$, and V represent the query, key, and value matrices, respectively, and $\mathrm{d_k}$ is the dimension of the key vector.

DenseNet-121 uses densely connected convolutional layers, where each layer receives inputs from all preceding layers, enhancing gradient flow and reducing parameters. EfficientNetB2 optimizes network depth, width, and resolution through compound scaling, improving efficiency and accuracy. ResNet50 introduces residual connections to address the vanishing gradient problem, allowing for deeper networks. VGG19 is characterized by its simplicity, consisting of deep convolutional layers followed by fully connected layers with small filters. The Vision Transformer (ViT) divides an image into small patches and learns their relationships using the Transformer mechanism. Swin Transformer employs a hierarchical architecture with sliding windows to effectively capture both local and global features. Data-efficient Image Transformers (DeiT) require less data for training while maintaining performance. Convolutional Vision Transformer (CvT) combines convolutional layers with Transformer encoders, enhancing spatial tokenization and local feature extraction through convolution operations.

To monitor the withdrawal speed during colonoscopy, we used perceptual hash functions to detect changes between consecutive video frames. The video processing workflow was implemented using OpenCV and PyTorch, and the perceptual hash (pHash) value for each frame was calculated using the imagehash library. The hash value calculation formula is given by: $H=\mathrm{pHasH}(I)$, where I represents the input image. By calculating the Hamming distance D between the hash values of adjacent frames, we quantify the visual changes as $D=H1-H2$, where $H1$ and $H2$ are the hash values of consecutive frames. This value indicates the extent of content change between frames. To clearly display the withdrawal speed, we overlaid a scale indicator on each video frame, with the position corresponding to the hash difference value D. We employed color coding: blue for normal speed ($D \leq 20$), yellow for warning speed ($21 \leq D \leq 30$), and red for hazardous speed ($D>30$).

Model interpretation

Despite the extensive application of advanced computer vision techniques in medical imaging, their widespread adoption in the medical community is still hampered by high computational costs, data constraints, and the black-box nature of deep learning. To enhance transparency, Explainable AI (XAI) has been introduced, aiming

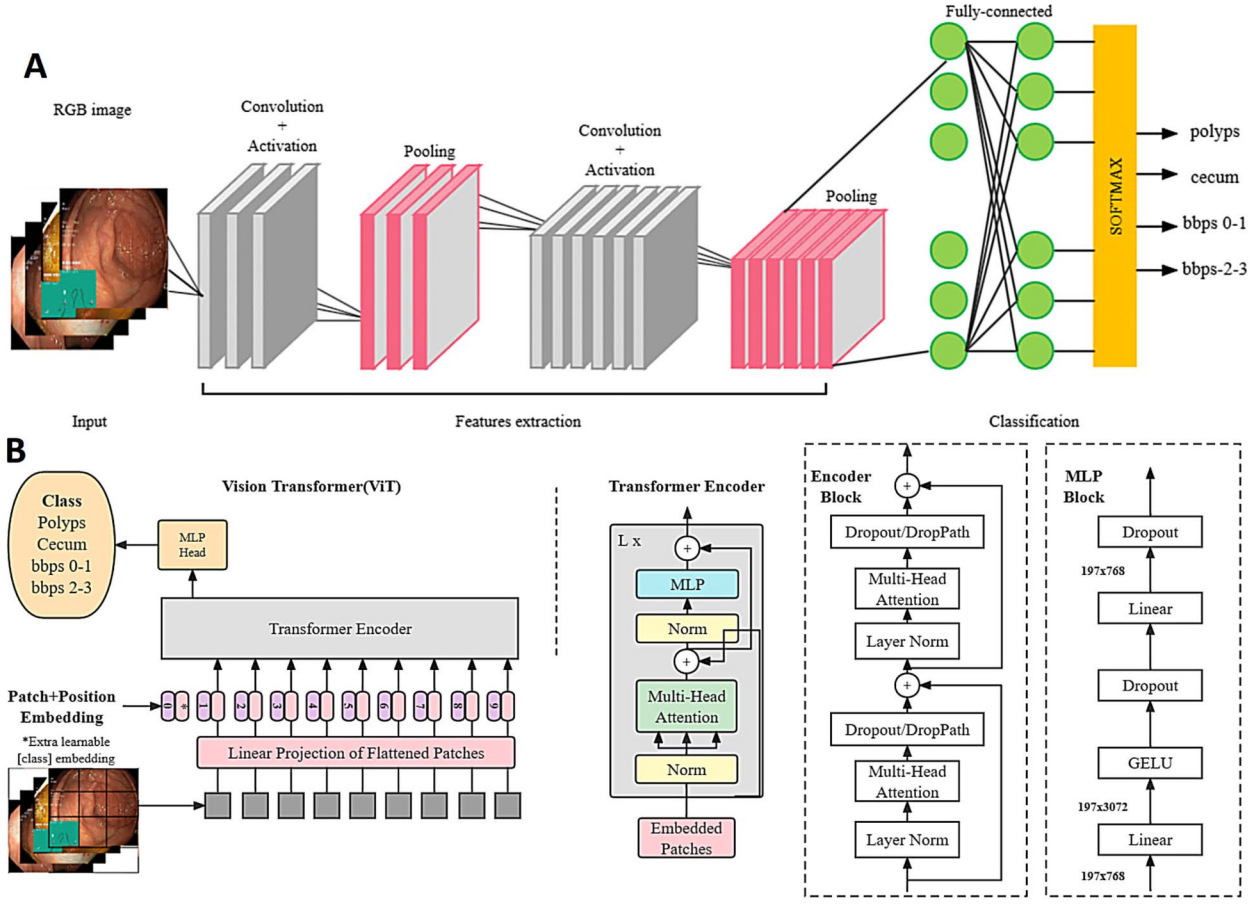

Fig. 3 (**A**) Schematic diagram of the CNN-based architecture. (**B**) Schematic diagram of the Transformer-based architecture, using VIT as an example

to elucidate the inner workings and decision-making processes of deep learning models. To combat this "black-box effect", we undertook an extensive explainability analysis of high-performance models based on CNN and Transformer architectures, employing techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM), Guided Grad-CAM, and SHAP [22–24]. Grad-CAM generates class-discriminative localization maps by using the gradients of any target concept (such as a specific class) flowing into the final convolutional layer to produce a coarse localization map highlighting important regions in the image. Mathematically, for a given class $_{\mathrm{c}}$, the gradient $\frac{\partial \mathrm{y}^{\mathrm{f}}}{\partial \mathrm{A}^{\mathrm{k}}}$ is computed with respect to feature map activations A^{k} of the last convolutional layer. These gradients are globally averaged to obtain weights $\alpha_{\mathrm{k}}^{\mathrm{c}}$, which are then used to compute the weighted sum of feature maps, followed by a ReLU operation: $\mathrm{L}_{\mathrm{Grad-CAM}}^{\mathrm{c}} = \mathrm{ReLU}\left(\sum_{\mathrm{k}} \alpha_{\mathrm{k}}^{\mathrm{c}} \mathrm{A}^{\mathrm{k}}\right)$.Guided Grad-CAM combines Grad-CAM and Guided Backpropagation, providing a more detailed visualization of how pixels influence decisions. SHAP values originate from cooperative game theory, providing a unified measure of feature importance. For image classification, SHAP assigns importance to each pixel, indicating its contribution to the prediction. SHAP values represent the difference in the expected model output when including versus excluding the feature, averaged over all possible feature combinations. This approach clarifies the role of each feature in the model's decision-making process. These methods collectively enhance our understanding of how the model interprets various colonoscopy images.

Deployment of Model Across multiple devices

To systematically enhance the quality of colonoscopy examinations, we developed a deep learning model and deployed it across multiple device endpoints, including desktop computers, laptops, and browsers in the endoscopy centers. This model aims to offer real-time quality control for colonoscopies, whether during or post-examination. Specifically, we initially acquired a PyTorch deep learning model tailored to our needs through transfer learning. Subsequently, to guarantee cross-platform deployment, we transitioned it into the ONNX format. Leveraging the ONNX Runtime, the model can efficiently operate across various operating systems (such as Linux, Windows, MacOS) and is optimized for different

hardware (like CPU, GPU). As an open standard for deep learning models, ONNX not only provides model interoperability but also presents us with a broad array of deployment options, ensuring the accuracy and efficiency of the colonoscopy examinations [25]. The model development and deployment process can be seen in Fig. 4.

Experimental platform and evaluation metrics

In this research, we employed a computing device equipped with an RTX 3060 graphics card (12GB VRAM), a CPU of 5×E5-2680 v4, and 350GB of disk space. With the aid of Python libraries such as TensorFlow (2.7.0), Keras (2.7.0), and OpenCV (4.5.4.60), we successfully constructed, trained, and executed image processing tasks with our deep learning model. For data organization, analysis, and visualization, we utilized tools like Pandas (1.3.4), NumPy (1.21.4), Matplotlib (3.5.0), and Plotly (5.4.0). Moreover, model optimization was accomplished using PyTorch (1.10.0+cu113), while saving and loading of the model were dependent on H5py (3.6.0).

This study employed a range of evaluation metrics to comprehensively assess the model's performance. The evaluation metrics include Area Under the Receiver Operating Characteristic Curve (AUC), recall, specificity, precision, accuracy, and F1 score. The calculation formulas are shown as Eq. (1) to (6).

(1) Recall or True Positive Rate (TPR):

$$Sensitivity = \frac{TP}{TP + FN}$$

(2) Specificity or True Negative Rate (TNR):

$$Specificity = \frac{TN}{TN + FP}$$

(3) Precision or Positive Predictive Value (PPV):

$$Precision = \frac{TP}{TP + FP}$$

(4)

$$\text{Accuracy} = \frac{TP + TN}{TP + \text{TN} + \text{FP} + \text{FN}}$$

(5) F1 Score$= 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity}$

(6) AUC: Area Under the Receiver Operating Characteristic Curve, measures the model's performance across different thresholds.

TP(True Positives)signifies the number of samples accurately identified as positive, TN (True Negatives) denotes the number of samples correctly identified as negative, FP (False Positives) refers to the number of samples erroneously predicted as positive, and FN (False Negatives) indicates the number of samples mistakenly predicted as negative.

Results

Performance comparison of various deep learning models on the validation set

This study conducted transfer learning fine-tuning based on pre-trained models from two major deep learning architectures: CNN and Transformer. In the CNN architecture, DenseNet-121, EfficientNetB2, ResNet50, and VGG19 were adopted, while in the Transformer architecture, ViT-Base-patch32-224, Swin-Small, DeiT-Small, and CvT-Small models were selected. The performance comparison of these models on the validation set is presented in Table 2.

In colonoscopy quality control tasks, EfficientNetB2 performed exceptionally well, achieving an accuracy of 0.992 on the test set, surpassing VGG19 (0.848) and Densenet121 (0.808). Furthermore, EfficientNetB2 excelled in precision, recall, and F1 score, registering 0.991, 0.989, and 0.990 respectively. In the Transformer architecture, the DeiT-Small model achieved an accuracy of 0.986, making it the best model within this architecture. Although its accuracy did not surpass that of EfficientNetB2, its performance remains noteworthy.

Prediction performance of the best model on the test set

To ensure the generalization performance of the model, we selected 358 colonoscopy images from Changshu Hospital Affiliated to Soochow University as an independent external test dataset for the best-performing model, EfficientNetB2. The advantage of using this independent test set lies in its ability to more accurately assess the model's performance in practical applications and to verify whether there are any overfitting issues.

The EfficientNetB2 model showcased superior colonoscopy image classification. Specifically, the AUC values for the BBPS 0–1 and BBPS 2–3 categories were an impressive 0.997 and 0.999, respectively, highlighting the model's superior discriminative capacity. The AUC for the cecum category stood at 0.996, while for the polyp category it was 0.993. On the whole, the model's average AUC, Precision, and Recall were 0.996, 0.948, and 0.952 respectively, all showcasing remarkable results, as depicted in Fig. 5A.

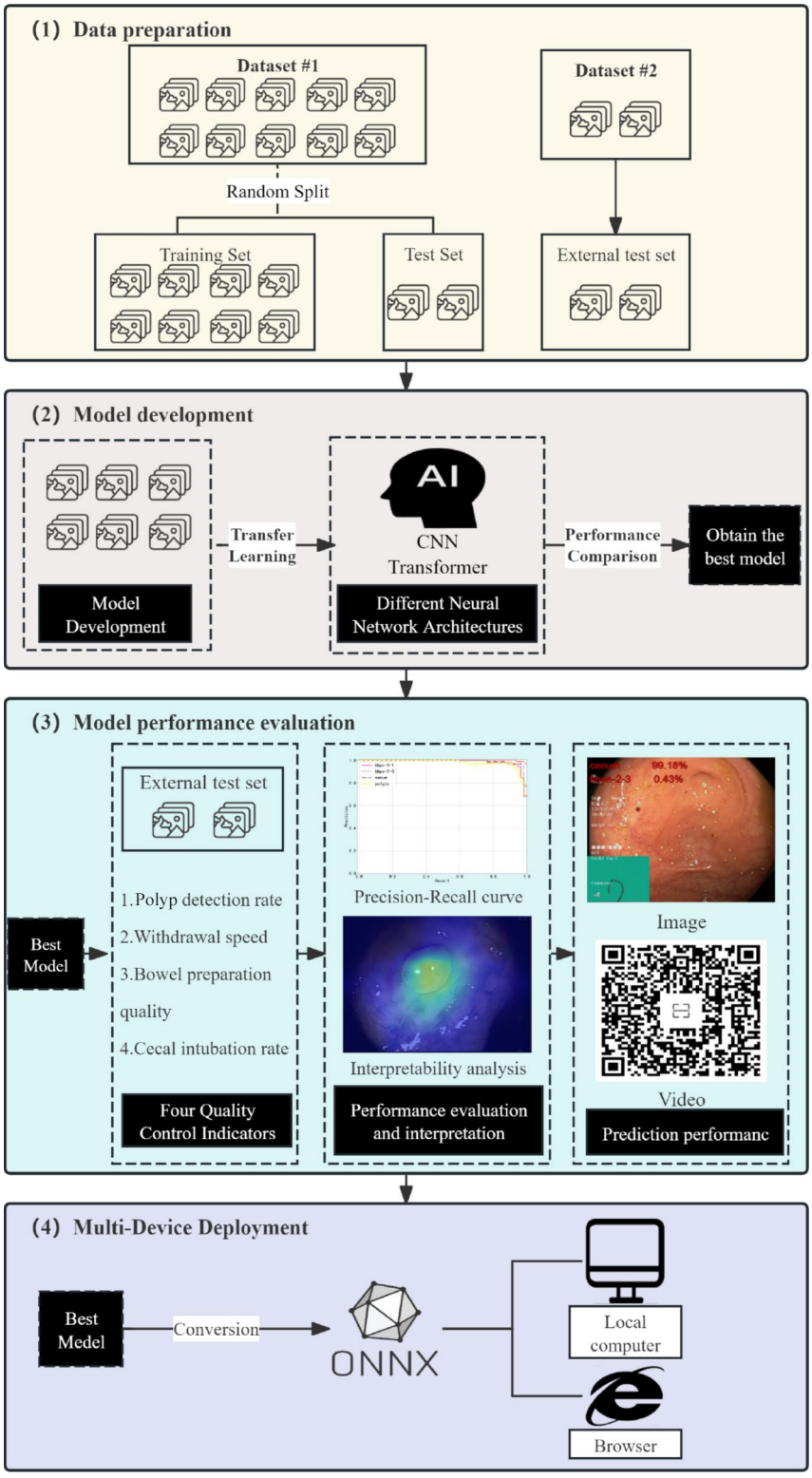

Fig. 4 Model development and deployment workflow; ONNX ensures model interoperability and diverse deployment options

Table 2 Performance comparison of different models on the validation set

Model Name	Accuracy	Precision	Recall	F1 score
DenseNet-121	0.991	0.989	0.988	0.989
EfficientNetB2	0.992	0.991	0.989	0.99
ResNet50	0.987	0.988	0.985	0.987
VGG19	0.991	0.991	0.989	0.99
Vit-Base-patch32-224	0.948	0.947	0.949	0.948
Swin-Small	0.908	0.906	0.913	0.907
CvT-Small	0.948	0.947	0.952	0.95
DeiT-Small	0.986	0.987	0.983	0.985

Figure 5B's PR curve highlights the model's near-optimal performance in Precision and Recall for the BBPS-0-1 and BBPS-2-3 categories. The cecum category's performance was also notably high, though slightly behind the first two categories. In contrast, for the Polyps category, while the Precision remained impressively high at 0.976, the Recall dropped to 0.890, suggesting the model might occasionally miss certain true positives. Nonetheless, the average precision (AP) for each category surpassed 0.97, further attesting to the model's robustness and consistency across various thresholds. Additionally, we conducted a confusion matrix analysis on the model's classification results, further confirming its accuracy and robustness across the different categories, with detailed outcomes presented in Fig. 5C.

In our experiments, despite the model's overall excellence, there were instances of misjudgments. Figure 5D displays images genuinely labeled as "cecum" but predicted by the model as "polyps." Similarly, in Fig. 5E, images truly labeled as "polyps" were misclassified as "cecum." These errors may arise from certain features in the images resembling polyps, leading to model confusion.

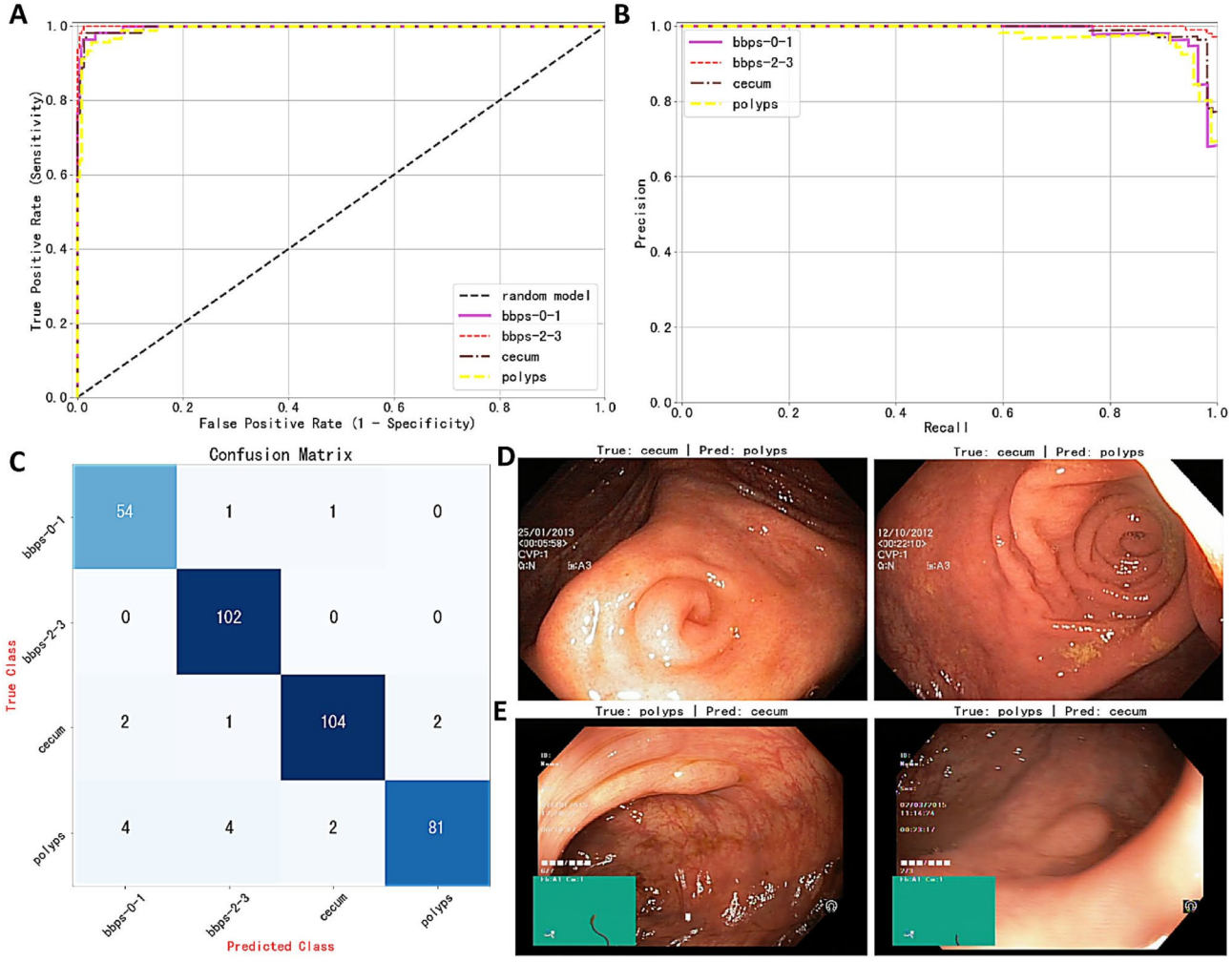

Fig. 5 Model performance on test set: (**A**) ROC curve; (**B**) PR curve; (**C**) Confusion matrix; (**D**) Image examples labeled as cesum but predicted as polyp by the model; (**E**) Image examples labeled as polyp but predicted as cecum by the model

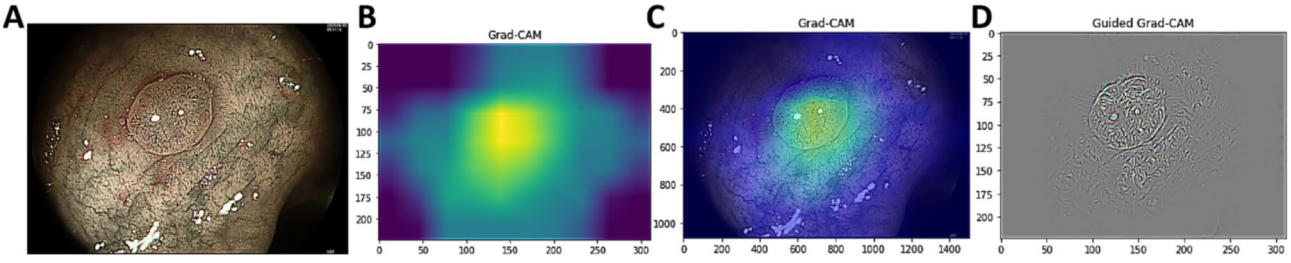

Fig. 6 Interpretation of the colonoscopy quality control model. (**A**) Original endoscopic image; (**B**) Pixel activation heatmap based on Grad-CAM; (**C**) Original image overlaid with activation heatmap; (**D**) Fine-grained heatmap from Guided Grad-CAM.

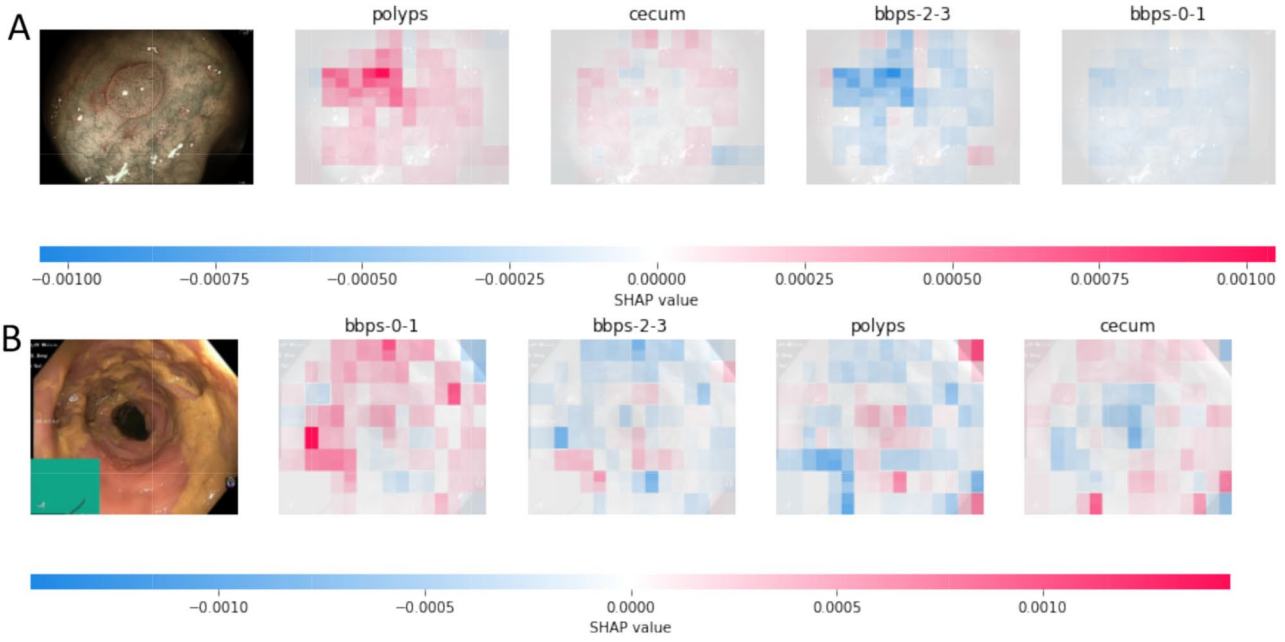

Fig. 7 SHAP interpretative analysis. (**A**) SHAP plot for a label identified as polyp with correct prediction; (**B**) SHAP plot for a label with BBPS score of 0–1 with correct prediction

Model interpretation

To gain a profound understanding of the decision-making mechanism of the colonoscopy quality control model, we employed the TorchCam library in conjunction with the Grad-CAM method for visual analysis. Figure 6A displays the original endoscopic image. Figure 6B presents the pixel activation heatmap derived from EfficientNetB2 feature extraction. These activations markedly delineate the image regions the model relies upon during its decision-making. Figure 6C superimposes the activation heatmap on the original image, where the yellow-green areas pinpoint the pivotal parts recognized by the model as polyps. To showcase the model's focal points in finer detail, Fig. 6D utilizes the Guided Grad-CAM technique, amalgamating both Grad-CAM and Guided Backpropagation, generating a heatmap that is both class-discriminative and granular, highlighting the intricate features the model depends on during classification.

To delve into the model's predictions, we utilized the SHAP (SHapley Additive exPlanations) method. As depicted in Fig. 7, Subfigure A and Subfigure B's actual categories are polyps and BBPS 0–1 score, respectively. The depth of color for each pixel in the figure signifies its influence on the prediction: red highlights positive contributions, while blue indicates negative ones. In Fig. 7A, the red region for the polyp category holds a distinct advantage over the cecum and the two BBPS categories, leading the model to accurately predict it as a polyp. Moreover, Fig. 7B is unequivocally identified as a BBPS 0–1 score.

Model-based video prediction and multi-terminal deployment

The PyTorch model with the best performance was converted to ONNX format and deployed on a local computer and web frontend, enabling real-time quality control for colonoscopy anywhere, anytime. Using the

OpenCV library, we captured each frame from the video source in real-time and fed them into the ONNX model for inference.

Figure 8 displays the prediction results for a single frame image. On the left side is the original image, with red text in the top left corner indicating the model's predictions for the first two categories along with their corresponding confidence levels. Correspondingly, the right side shows a bar chart representation of the confidence levels for each category. Figure 8A and B respectively show the model's predictions and confidence levels for cecum images and polyp images. Figure 8C and D present the model's real-time prediction results on video in the form of QR codes; users can scan the QR code with their phones to watch the corresponding videos.

Figure 8C demonstrates the prediction results after the model's deployment on a local computer. The scale at the bottom left of the video screen displays the current withdrawal speed in real time. When the withdrawal speed is too fast, the scale falls into the red area and displays a "hazardous speed" warning; when the speed is within a reasonable range, it displays a blue "normal speed" indication; otherwise, it shows a yellow "warning speed" alert. Figure 8D presents the real-time prediction results using the local computer's camera.

Discussion

This study developed eight deep learning models for performing colonoscopy quality control tasks, including four CNN architectures and four Transformer architectures. From the largest gastrointestinal endoscopy dataset, HyperKvasir, we selected 3,831 images for model training and tested the performance on an independent test set. The EfficientNetB2 model performed the best among all models. By incorporating real-time withdrawal speed monitoring, this AI system integrated four key quality control indicators related to colonoscopy. This model has been successfully deployed across multiple platforms, enabling real-time video prediction. This study is the first to compare the performance of CNN and Transformer architectures in colonoscopy quality control and to identify the optimal model.

Colorectal cancer's high incidence and mortality rates in China present significant challenges to public health and the economy [26]. Timely diagnosis through colonoscopy is critical for improving patient outcomes. However, the substantial rate of missed diagnoses, particularly for early-stage tumors, necessitates enhanced quality control measures for colonoscopies. Concurrently, heightened health awareness has led to increased demands for colonoscopic exams, placing additional pressures on endoscopic services and escalating both costs and time. The expansion of Artificial Intelligence (AI) in healthcare, especially through deep learning, has provided new avenues for efficient data and image processing, potentially

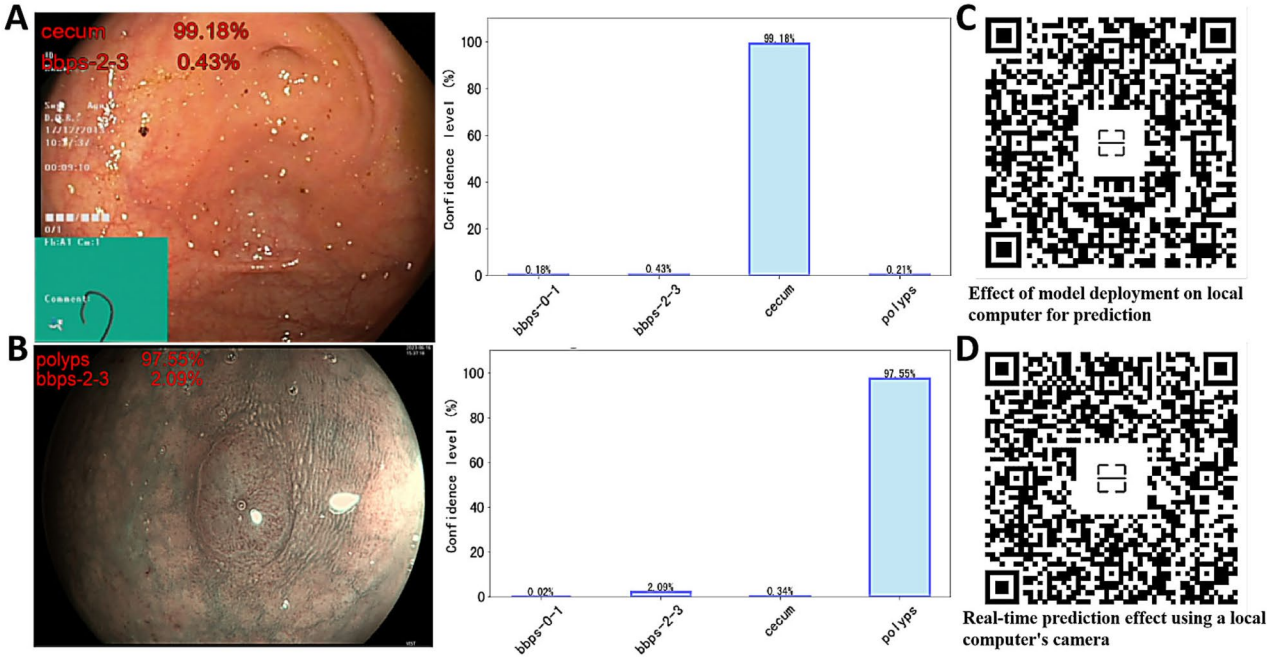

Fig. 8 Model deployment for single-frame images and video predictions with confidence levels. (**A**) A single-frame image of the cecum. (**B**) A single-frame image of a polyp. (**C**) Real-time prediction performance on a local computer setup. (**D**) Real-time prediction using a local computer camera

elevating the quality control of digestive endoscopy. Professional societies such as the European Society of Gastrointestinal Endoscopy, the Digestive Endoscopy Branch of the Chinese Medical Association, and the American Society for Gastrointestinal Endoscopy have underscored four critical quality indicators for colonoscopy: polyp detection, cecal intubation rate, withdrawal speed, and bowel preparation quality [12–14]. Given the central importance of these four indicators in assessing the quality of colonoscopy, this study trained a deep learning classification model designed to comprehensively cover these crucial aspects.

In a recent study by Yao et al. [27], the team used convolutional neural networks to develop an endoscopic quality control system called Endo.Adm. This system significantly improved detection rates for adenomas and early-stage gastric cancers, demonstrating the potential of deep learning in enhancing colonoscopy quality control. However, Endo.Adm's exclusion of the Transformer architecture and its opaque decision-making process warrant further scrutiny. Our study explored this by integrating Vision Transformer (ViT), Swin Transformer, DeiT, and Convolutional Vision Transformer (CvT) models into the Transformer framework. The DeiT model, outperformed others, achieving an accuracy of 0.986, an impressive feat even though it did not eclipse EfficientNetB2's results.

We conducted a thorough interpretability analysis on the highest-performing models using techniques such as Grad-CAM, Guided Grad-CAM, and SHAP. Through the torchcam library, Grad-CAM elucidated the decision-making process by visualizing activation hotspots, pinpointing areas critical to the model's assessments. The Guided Grad-CAM refined this visualization, producing high-resolution, class-discriminative heatmaps, and SHAP analysis quantified the predictive contribution of individual pixels. Collectively, these methods provided an insightful elucidation of the model's decision-making processes.

In the colonoscopy quality control task, EfficientNetB2 was rated as the best-performing model compared to the other seven models. It demonstrated excellent performance across all categories, with F1 scores reaching or exceeding 93%. This reflects the model's high precision and recall, indicative of a performance equilibrium. Moreover, we transitioned EfficientNetB2 to ONNX format, facilitating deployment on diverse devices. In real-time processing of camera and video inputs, the model maintained exemplary classification accuracy with efficiency exceeding 60 frames per second, assuring prompt feedback for real-time applications. This high-efficiency profile of EfficientNetB2 is attributable to its balanced scaling strategy across network depth, width, and resolution, which optimizes for rapid inference while preserving accuracy.

Currently favored international standards for bowel preparation assessment are the Boston Bowel Preparation Scale (BBPS) and the Ottawa Bowel Preparation Scale (OBPS), with our study employing the former. Continual reliance on these scales poses a challenge in clinical settings, notably due to assessment inconsistencies and subjective biases inherent to endoscopic practitioners. Disparities in bowel cleanliness evaluation among medical staff highlight the necessity for an objective and streamlined assessment method. The EfficientNetB2 model demonstrated exceptional performance in identifying bowel cleanliness on the validation set, particularly in the BBPS (0–1) and BBPS (2–3) categories, with AUC values of 0.997 and 0.999, respectively, indicating a high level of discrimination ability. On the test set, the model's AUC values reached 0.986 and 0.997, further confirming its superior performance.

Previous studies have often focused on developing AI models with a single function, such as withdrawal speed monitoring [28], bowel cleanliness assessment [29], and polyp detection models [30]. However, colonoscopy quality control is a comprehensive evaluation system, and single-function models are clearly insufficient to meet the needs. Moreover, maintaining an appropriate withdrawal speed helps improve polyp detection rates, indicating that these quality indicators are not independent but rather complementary. Therefore, developing a multifunctional AI-assisted system holds promise for more proactively enhancing the quality of colonoscopy.

Despite providing new insights into the application of deep learning in colonoscopy quality control, this study faces several challenges. First, we plan to prospectively include images and video data from more medical centers to further test the model's performance in real clinical settings, thereby making our findings more broadly representative. Second, we intend to conduct human-machine comparison experiments, comparing the model with endoscopists of varying levels of experience. This systematic evaluation will provide valuable data and insights for technological improvements and clinical applications.

Conclusions

This study covers the entire process from model training, validation, testing, interpretability analysis, to terminal deployment, resulting in the development of a comprehensive AI-assisted system. This system integrates four key quality control indicators for colonoscopy: real-time monitoring of withdrawal speed, enhanced polyp detection rate, automatic assessment of bowel preparation quality, and cecal intubation rate. Through this single

model, comprehensive management and improvement of colonoscopy quality are achieved.

Acknowledgements
The authors thank the team behind the HyperKvasir dataset for providing the data used in this study.

Author contributions
CJ and XXD worked on the study design. WGH and ZJZ worked on data collection. ZZH, XKJ, DY, and CJ worked on data analysis. DY worked on manuscript preparation. XXD, CJ, and XKJ provided administrative, technical, or material support. XXD supervised the study. All authors have made a significant contribution to this study and have approved the final manuscript.

Funding
Changshu City Science and Technology Plan Project: CS202116. Health Informatics Key Support Discipline Funding of Suzhou City: SZFCXK202147. No funding body had any role in the design of the study and collection, analysis, interpretation of data, or in writing the manuscript.

Data availability
The datasets analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
This study has been approved by the Ethics Review Committee of Changshu Hospital Affiliated to Soochow University (Approval No.: L20230930). This study was performed in accordance with the Declaration of Helsinki, and written informed consent was obtained from all participants.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 3 November 2023 / Accepted: 6 August 2024
Published online: 09 August 2024

References

1. Gunter MJ, Alhomoud S, Arnold M, Brenner H, Burn J, Casey G, Chan AT, Cross AJ, Giovannucci E, Hoover R, et al. Meeting report from the joint IARC–NCI international cancer seminar series: a focus on colorectal cancer. Ann Oncol. 2019;30(4):510–9.
2. Gupta S, Lieberman D, Anderson JC, Burke CA, Dominitz JA, Kaltenbach T, Robertson DJ, Shaukat A, Syngal S, Rex DK. Recommendations for Follow-Up after Colonoscopy and Polypectomy: a Consensus Update by the US Multi-society Task Force on Colorectal Cancer. Am J Gastroenterol. 2020;115(3):415–34.
3. Bretthauer M, Kalager M, Adami H. Do's and don'ts in evaluation of endoscopic screening for gastrointestinal cancers. Endoscopy. 2016;48(1):75–80.
4. Li Y, Zhu Z, Chen JJ, Jing JC, Sun C, Kim S, Chung P, Chen Z. Multimodal Endoscopy for colorectal cancer detection by optical coherence tomography and near-infrared fluorescence imaging. Biomed Opt Express. 2019;10(5):2419–29.
5. Pop OL, Vodnar DC, Diaconeasa Z, Istrati M, Bințințan A, Bințințan VV, Suharoschi R, Gabbianelli R. An overview of gut microbiota and Colon diseases with a focus on adenomatous Colon polyps. Int J Mol Sci 2020, 21(19).
6. Su H, Lao Y, Wu J, Liu H, Wang C, Liu K, Wei N, Lin W, Jiang G, Tai W, et al. Personal instruction for patients before colonoscopies could improve bowel preparation quality and increase detection of colorectal adenomas. Ann Palliat Med. 2020;9(2):420–7.
7. Gómez-Reyes E, Tepox-Padrón A, Cano-Manrique G, Vilchis-Valadez NJ, Mora-Bulnes S, Medrano-Duarte G, Chaires-Garza LG, Grajales-Figueroa G, Ruiz-Romero D. Téllez-Ávila FI: a low-residue diet before colonoscopy tends to improve tolerability by patients with no differences in preparation quality: a randomized trial. Surg Endosc. 2020;34(7):3037–42.
8. Belderbos TDG, Grobbee EJ, van Oijen MGH, Meijssen MAC, Ouwendijk RJT, Tang TJ, ter Borg F, van der Schaar P, Le Fèvre DM, Stouten MT, et al. Comparison of cecal intubation and adenoma detection between hospitals can provide incentives to improve quality of colonoscopy. Endoscopy. 2015;47(8):703–9.
9. Zhang Q, Dong Z, Jiang Y, Zhan T, Wang J, Xu S. The Impact of Sedation on Adenoma Detection Rate and Cecal Intubation Rate in. *Gastroent Res Pract* 2020, 2020:3089094.
10. Aziz M, Haghbin H, Gangwani MK, Nawras M, Nawras Y, Dahiya DS, Sohail AH, Lee-Smith W, Kamal F, Shaukat A. 9-Minute Withdrawal Time improves Adenoma Detection Rate compared with 6-Minute Withdrawal Time during Colonoscopy: a Meta-analysis of Randomized controlled trials. J Clin Gastroenterol. 2023;57(9):863–70.
11. Yamaguchi H, Fukuzawa M, Minami H, Ichimiya T, Takahashi H, Matsue Y, Honjo M, Hirayama Y, Nutahara D, Taira J, et al. The relationship between Post-colonoscopy Colorectal Cancer and Quality indicators of Colonoscopy: the latest single-center Cohort Study with a review of the literature. Intern Med (Tokyo Japan). 2020;59(12):1481–8.
12. Rembacken B, Hassan C, Riemann JF, Chilton A, Rutter M, Dumonceau J, Omar M, Ponchon T. Quality in screening colonoscopy: position statement of the European Society of Gastrointestinal Endoscopy (ESGE). Endoscopy. 2012;44(10):957.
13. Rex DK, Schoenfeld PS, Cohen J, Pike IM, Adler DG, Fennerty MB, Lieb JGN, Park WG, Rizk MK, Sawhney MS, et al. Quality indicators for colonoscopy. Am J Gastroenterol. 2015;110(1):72–90.
14. Committee of Colorectal Cancer Quality Control, National Cancer Center and National Clinical Research Center for Cancer. Guidelines for the standardized diagnosis and treatment quality control indicators of primary colorectal cancer (2022 Edition). Chin J Oncol. 2022;44(7):623–7. (In Chinese).
15. Zhao S, Yang W, Wang S, Pan P, Wang R, Chang X, Sun Z, Fu X, Shang H, Wu J, et al. Establishment and validation of a computer-assisted colonic polyp localization system based on deep learning. World J Gastroentero. 2021;27(31):5232–46.
16. Zhou W, Yao L, Wu H, Zheng B, Hu S, Zhang L, Li X, He C, Wang Z, Li Y, et al. Multi-step validation of a deep learning-based system for the quantification of bowel preparation: a prospective, observational study. Lancet Digit Health. 2021;3(11):e697–706.
17. Liu Z, Lv Q, Yang Z, Li Y, Lee CH, Shen L. Recent progress in transformer-based medical image analysis. Comput Biol Med. 2023;164:107268.
18. Karaman A, Pacal I, Basturk A, Akay B, Nalbantoglu U, Coskun S, Sahin O, Karaboga D. Robust real-time polyp detection system design based on YOLO algorithms by optimizing activation functions and hyper-parameters with artificial bee colony (ABC). Expert Syst Appl. 2023;221:119741.
19. Gong R, Yao L, Zhang L, Li X, Zhang J, Li J, Jiang X, Zhao Y, Wang J, Zhang C, et al. Complementary effect of the proportion of Overspeed frames of Withdrawal and Withdrawal Time on reflecting Colonoscopy Quality: a retrospective, observational study. Clin Transl Gastroen. 2023;14(3):e566.
20. Ahmad OF. Deep learning for automated bowel preparation assessment during colonoscopy: time to embrace a new approach? In., 3; 2021: e685–6.
21. Borgli H, Thambawita V, Smedsrud PH, Hicks S, Jha D, Eskeland SL, Randel KR, Pogorelov K, Lux M, Nguyen DTD et al. HyperKvasir, a comprehensive multiclass image and video dataset for gastrointestinal endoscopy. Sci Data 2020, 7(1).
22. 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.
23. 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.
24. Ye T, Li S, Zhang Y. Genomic pan-cancer classification using image-based deep learning. Comput Struct Biotec. 2021;19:835–46.
25. Li P, Wang X, Huang K, Huang Y, Li S, Iqbal M. Multi-model running latency optimization in an Edge Computing paradigm. Sensors 2022, 22(16).
26. Xu L, Zhao J, Li Z, Sun J, Lu Y, Zhang R, Zhu Y, Ding K, Rudan I, Theodoratou E, et al. National and subnational incidence, mortality and associated factors of colorectal cancer in China: a systematic analysis and modelling study. J Glob Health. 2023;13:4096.

27. Yao L, Liu J, Wu L, Zhang L, Hu X, Liu J, Lu Z, Gong D, An P, Zhang J, et al. A gastrointestinal Endoscopy Quality Control System Incorporated with Deep Learning Improved Endoscopist performance in a Pretest and Post-test Trial. Clin Transl Gastroen. 2021;12(6):e366.
28. Lui TKL, Ko MKL, Liu JJ, Xiao X, Leung WK. Artificial intelligence–assisted real-time monitoring of effective withdrawal time during colonoscopy: a novel quality marker of colonoscopy. Gastrointest Endosc. 2024;99(3):419–27.
29. Wang Y, Jheng Y, Sung K, Lin H, Hsin I, Chen P, Chu Y, Lu D, Wang Y, Hou M et al. Use of U-Net Convolutional neural networks for automated segmentation of fecal material for objective evaluation of Bowel Preparation Quality in Colonoscopy. Diagnostics (Basel Switzerland) 2022, 12(3).
30. Pacal I, Karaman A, Karaboga D, Akay B, Basturk A, Nalbantoglu U, Coskun S. An efficient real-time colonic polyp detection with YOLO algorithms trained by using negative samples and large datasets. Comput Biol Med. 2022;141:105031.

Publisher's Note