Journal of Computer Science

Original Research Paper

AI-Guided Anatomical Landmark and Abnormality Detection for Autonomous Endoscopy Examination

[1,2]Md Shakhawat Hossain, [2]Munim Ahmed, [2]Md Sahilur Rahman, [3]Mahreen Tabassum, [3]Fariha Karim, [4]Md Aulad Hossain, [1,2]Razib Hayat Khan, [1,2]M. M. Mahbubul Syeed and [1,2]Mohammad Faisal Uddin

[1]*Department of Computer Science and Engineering, Independent University, Bangladesh, Dhaka, Bangladesh*
[2]*RIoT Research Center, Independent University, Bangladesh, Dhaka, Bangladesh*
[3]*Department of Computer Science and Engineering, American International University-Bangladesh, Dhaka, Bangladesh*
[4]*Department of Gastroenterology, Bangabandhu Sheikh Mujib Medical University Hospital, Dhaka, Bangladesh*

Article history
Received: 22-04-2024
Revised: 22-05-2024
Accepted: 24-05-2024

Corresponding Author:
Md Shakhawat Hossain
Department of Computer Science and Engineering, Independent University, Bangladesh, Dhaka, Bangladesh
Email: shakhawat@iub.edu.bd

Abstract: Endoscopy is the routine medical procedure to observe tumors in the human Gastrointestinal (GI) tract by inserting an endoscope, a thin, flexible, tube-like instrument with a light source and camera. Traditionally, an endoscopist performs the endoscopy, orients the endoscope within these structures and navigates this through the help of familiar anatomical landmarks to reach the abnormalities and mark them. Identifying landmarks and abnormalities is critical for the maneuver and the success of endoscopy, which is related to the patient's comfort, injury and accurate diagnosis. The manual naked-eye-observation maneuver and examination are highly challenging, take a long time and often cause discomfort to the patients and the endoscopists. As a result, several AI-based landmark detection methods have been proposed recently to facilitate autonomous endoscopy examination. However, these methods lack accuracy and consider only limited landmarks. This study presents a Data-efficient image Transformer (DeiT)-based method to detect anatomical landmarks and anomalies for autonomous endoscopy. The proposed method detected 23 landmarks and anomalies from the entire GI tract with 99% accuracy and precision, outperforming the state-of-the-art (91%). Moreover, this method took only 0.045 sec to identify a landmark. The phi coefficient (0.997) indicated a strong positive association between the proposed method and clinical ground truth. Strong association, high accuracy and rapid speed ensured the reliability of the proposed method for autonomous endoscopy examination.

Keywords: Endoscopy, Anatomical Landmarks, Transformer, Abnormality Detection, Computer Aided Diagnosis

Introduction

Over the past few decades, there has been a notable increase in Gastrointestinal (GI) related health issues worldwide. The World Health Organization (WHO) reports that colorectal and stomach cancers currently account for the majority of cancer cases worldwide, with 2.8 million new cases and 1.6 million deaths from these cancers in 2018. GI tract-related cancers have a combined mortality of about 63% with 2.2 million deaths per year (Ferlay *et al*., 2015). The primary step in diagnosing and treating Gastrointestinal (GI) problems is endoscopy to examine and diagnose abnormalities. An endoscopist performs an endoscopy by inserting the endoscope into the suspected GI tract region through a minor incision or natural body opening, such as the mouth, anus, or urethra. Then, the camera attached to the endoscope sends the video signal, which endoscopists observe on a screen to examine in real time. Endoscopists use anatomical landmarks such as Z-line, pylorus, cecum, ileum and others inside the GI tract observed on the screen to safely navigate the endoscope to the region of interest during endoscopy. Anatomical landmarks are also used to confirm the location of a lesion and help follow a predetermined path for endoscopy. Endoscopists must also identify abnormalities and pathological findings such as polyps and hemorrhoids. Some findings are used to indicate the quality of views inside the GI tract. For example, the Boston Bowel Preparation Scale (BBPS), used to indicate the degree of clean bowel or the quality of mucosal views, is necessary for accurate endoscopy. Some pathological findings include therapeutic interventions that show previously treated regions or treatment markings. For example, polyps are lifted with submucosal injection using a solution before resectioning. This finding is termed dyed lifted polyps, which appear blue in endoscopy. After the resection, the resection margins and site also appear blue, termed dyed resection margin. There are instances when the endoscopists also need to cut tissue or dilate narrow passages. The landmarks, abnormalities and pathological findings can be observed in entire GI tracts, which are classified according to the minimal standard terminology by the World Endoscopy Organization (Aabakken *et al*., 2014). This terminology provided a standardized model for categorizing the endoscopic findings. The success of the endoscopy and the comfort and safety of the patient depends on accurately identifying landmark and pathological findings and abnormalities. In this research, we have followed the guidelines of the world endoscopy organization (Aabakken *et al*., 2014) and designed the proposed system to classify them into 23 classes, which include major anatomical landmarks, abnormalities, pathological findings, mucosal views and therapeutic interventions, as shown in Fig. 1. However, for simplicity, we termed them as anatomical landmarks and abnormalities.

At the moment, endoscopists use their naked eyes to identify landmarks and abnormalities manually. The highly challenging manual examination takes a lot of time and effort, frequently resulting in discomfort and occasionally causing injury. This manual examination-based endoscopy depends on an expert's availability and causes delays in healthcare, particularly in Bangladesh, where there is a shortage of healthcare professionals (Hossain *et al*., 2022). Furthermore, our investigation revealed interobserver variability in the manual examination. In this study, three endoscopists were given 2272 endoscopy images and asked to classify them into one of the 23 classes. The results are presented in Table 1, which shows inter-observer variation. In another study, Kaminski *et al.* (2010) reported a 20% polyps miss-rate in the colon (Kaminski *et al*., 2010). These findings suggest the need for an automated endoscopy examination method that can improve the accuracy of endoscopy examination and diagnosis to reduce morbidity and deaths associated with GI disease. Artificial Intelligence (AI) has recently shown significant success in developing image-based automated analysis and diagnosis systems, which have aided medical professionals in delivering high-quality care to many patients in big hospitals and laboratories. Inspired by these studies, we proposed an AI-based solution for autonomous endoscopy examination in this study.

Table 1: Comparison of proposed method with experts manual examination

Anatomical landmarks and abnormalities	Examination by expert A %	Examination by expert B %	Examination by expert C %	Classification by proposed method %
Barrett	78/86 (90)	82/86 (95)	86/86 (100)	86/86 (100)
Barrett's short segment	92/99 (93)	97/99 (98)	99/99 (100)	99/99 (100)
BBPS 0-1	89/100 (89)	92/100 (92)	96/100 (96)	100/100 (100)
BBPS 2-3	66/77 (85)	71/77 (92)	72/77 (93)	76/77 (98)
Cecum	107/116 (92)	110/116 (95)	115/116 (99)	115/116 (99)
Polyps	107/107 (100)	107/107 (100)	107/107 (100)	106/107 (99)
Dyed lifted polyps	97/97 (100)	97/97 (100)	97/97 (100)	97/97 (100)
Dyed resection margin	93/93 (100)	93/93 (100)	93/93 (100)	91/93 (97)
Esophagitis-a	88/99 (88)	91/99 (91)	94/99 (94)	99/99 (100)
Esophagitis b-d	96/107 (89)	97/107 (90)	99/107 (92)	106/107 (99)
Hemorrhoids	102/102 (100)	102/102 (100)	102/102 (100)	102/102 (100)
Ileum	93/101 (92)	95/101 (94)	98/101 (97)	101/101 (100)
Impacted stool	96/96 (100)	96/96 (100)	96/96 (100)	96/96 (100)
Pylorus	88/102 (86)	92/102 (90)	95/102 (93)	102/102 (100)
Retroflex rectum	98/100 (98)	99/100 (99)	99/100 (99)	100/100 (100)
Retroflex stomach	101/112 (90)	105/112 (94)	107/112 (95)	112/112 (100)
Ulcerative colitis 0-1	72/81 (88)	78/81 (96)	78/81 (99)	81/81 (100)
Ulcerative colitis 1-2	87/107 (81)	88/107 (82)	90/107 (84)	107/107 (100)
Ulcerative colitis 2-3	101/110 (92)	105/110 (95)	104/110 (95)	110/110 (100)
Ulcerative colitis 1	77/97 (79)	79/97 (81)	82/97 (84)	97/97 (100)
Ulcerative colitis 2	72/78 (92)	71/78 (91)	73/78 (93)	78/78 (100)
Ulcerative colitis 3	96/104 (92)	93/104 (89)	96/104 (92)	104/104 (100)
Z-line	88/101 (87)	93/101 (92)	97/101 (96)	101/101 (100)
Overall	2084/ 2272 (91)	2133/2272 (93)	2175/2272 (95)	2266/2272 (99)

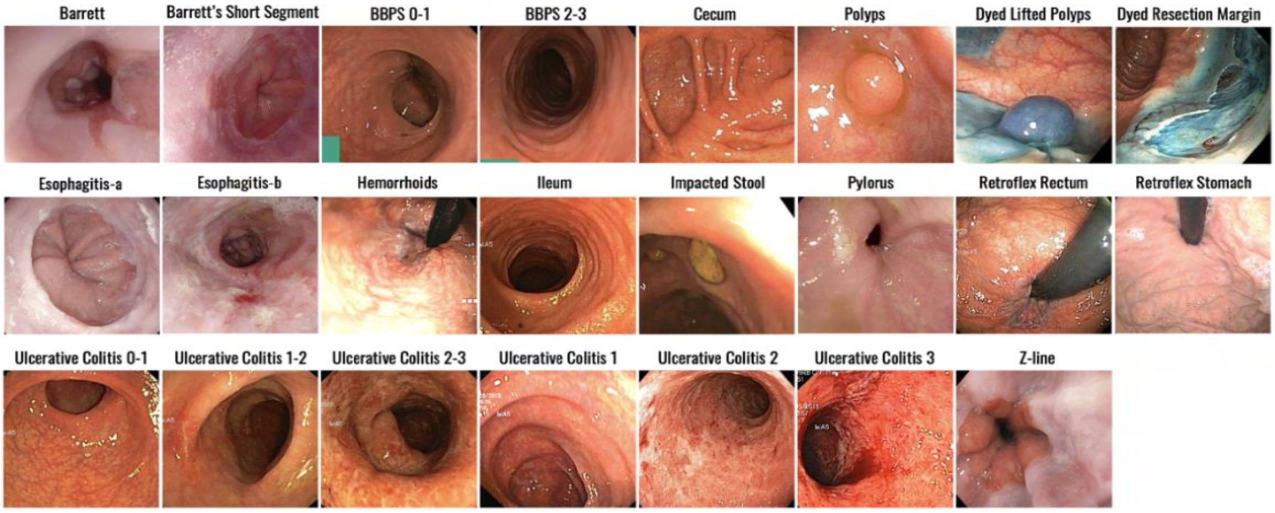

Fig. 1: Major anatomical landmarks and abnormalities of GI tract

Previously, several AI-assisted methods were proposed for this purpose; however, they failed to meet the primary requirements for the automation of endoscopy examination for practical use, which are accuracy, speed and ability to identify diverse landmarks. Most of the methods detected only selected abnormalities or conditions such as polyps. Other methods detected limited landmarks or abnormalities from specific parts of the GI tract, such as the upper or lower GI tract. These methods are suitable for the diagnosis of a specific condition (Hossain *et al*., 2023; Ozawa *et al*., 2020; Bour *et al*., 2019; Tomar *et al*., 2021; Misawa *et al*., 2021; Aliyi *et al*., 2023) (i.e., hyperproliferation, severe dysplasia) or cancer (Suzuki *et al*., 2021; Luo *et al*., 2019; Hirasawa *et al*., 2018; Iwagami *et al*., 2021) (i.e., adenocarcinoma, colorectal cancer, stomach cancer), but not suitable for fully-automated endoscopy examination. Only a few methods were intended for complete autonomous endoscopy examination; however, they detected only limited landmarks or abnormalities, additionally failed to achieve adequate accuracy and practical usability (Che *et al*., 2021; Tran *et al*., 2021; Ayyoubi Nezhad *et al*., 2022; Borgli *et al*., 2020). Che *et al*. (2021) proposed a method for detecting anatomical landmarks in the lower GI tract from colonoscopy videos. They trained ResNet-101, a CNN-based network, to detect three landmarks, which achieved 92% accuracy. Tran *et al*. (2021) proposed another CNN-based method for detecting anatomical landmarks from the upper GI tract. They have detected ten landmarks with 97% accuracy, suitable automated esophagogastroduodenoscopy. These methods Che *et al*. (2021); Tran *et al*. (2021) were trained and tested using locally collected data. Ayyoubi Nezhad *et al*. (2022) proposed a method for detecting landmarks from the entire GI tract. This method also relied on the CNN-based network to detect three landmarks with 99% accuracy. This method was trained on the Kvasir dataset (Pogorelov *et al*., 2017), a public dataset containing ten anatomical landmarks and abnormalities of the GI tract. Despite being developed for the entire GI tract, this method only considered three landmarks, making it impractical; also, the CNN model was trained for a limited number of epochs, resulting in an over-fitted network. Borgli *et al*. (2020) prepared a dataset called hyper kvasir to facilitate the development of autonomous endoscopy examination methods and proposed a CNN-based method for landmark and abnormality detection that works for the entire GI tract. Hyper kvasir is currently the largest dataset with the most landmarks and abnormalities. Borgli *et al*. (2020) combined two CNN-based models, ResNet-152 and DenseNet-161, to predict the final class. However, this method only achieved an average accuracy of 91% for detecting 23 landmarks. Even though its accuracy is insufficient, this method was developed to identify landmarks and abnormalities from the entire GI system and detected diverse landmarks of 23 types, which aligns with the goal of this study.

All the abovementioned methods considered only a few landmarks for detection, except Borgli Hanna's method (Borgli *et al*., 2020). Moreover, none of these methods underwent an analysis to determine their feasibility for practical use. Thus, these systems are not suitable for complete autonomous endoscopy examination. This study aims to develop a method for autonomous endoscopy examination for the entire GI tract. The major contributions of this study include (1) The development of an AI-assisted landmark and abnormality classification method for autonomous endoscopy examination, (2) A comparison of CNN and transformer-based networks for endoscopic landmark classification with limited data and (3) Feasibility analysis of the proposed method for practical use.

Methods

Data Collection, Evaluation and Balancing

Several datasets of the GI tract are available; however, most of these datasets have limited landmarks or abnormality images. Most of them only contained polyps images, the most common abnormality. Hyper kvasir (Borgli *et al*., 2020) is the largest dataset with the most landmarks suitable for developing an AI-assisted autonomous system. Therefore, we have used and enhanced the hyper kvasir dataset for our experiment. The images of hyper kvasir were collected from routine clinical examinations performed at the department of gastroenterology of Baerum hospital, Vestre Viken hospital trust, Norway, from 2008-2016.

The endoscopy was performed using standard Olympus and Pentax endoscopy machines. The images were extracted from the endoscopy videos and labeled by experienced gastrointestinal endoscopists.

As illustrated in Fig. 2, the hyper kvasir dataset had 10,662 labeled images of 23 different types, which included anatomical landmarks, pathogenic findings, therapeutic actions and typical abnormalities. The images manifest findings from the lower and upper GI tracts, including three anatomical landmarks from each GI tract, four pathological findings from the upper and eight from the lower GI and two therapeutic interventions and three mucosal landmarks from the lower GI tract. Upper GI tract anatomical landmarks include the pylorus, z-line and retroflex stomach. Barretts, Barrett's short segment, esophagitis a and esophagitis b-d are pathological findings of the upper GI tract. The cecum, ileum and retroflex rectum are the anatomical landmarks of the lower GI tract. Polyps, hemorrhoids, ulcerative colitis 1, ulcerative colitis 2, ulcerative colitis 3, ulcerative colitis 0-1, ulcerative colitis 1-2 and ulcerative colitis 2-3 are pathological findings of lower GI. Lower GI therapeutic interventions include dyed lifted polyps and resection margins, whereas mucosal landmarks include BBPS 0-1, BBPS 2-3 and impacted stool. For simplicity, we have termed these as anatomical landmarks and abnormalities in this study.

The number of images per class in the hyper kvasir dataset is not balanced; also, just a few images are available for several classes such as Ileum, hemorrhoids, ulcerative colitis 0-1, ulcerative colitis 1-2, ulcerative colitis 2-3 and Barretts. It is challenging to train AI models with skewed data. Furthermore, some classes have fewer than ten images, making balancing the dataset using augmentation approaches impractical. Over-fitting occurs when a model is trained with augmented images generated from a small set. As a result, in this study, we initially enhanced the quantity of images for minority classes by including images from a local hospital. The images were captured using a Pentax endoscopy machine. The images collected from the local hospital were included only in the test dataset to evaluate the robustness of the models. Following that, each class had at least 200 images. The dataset was then augmented by flipping and rotating the images. In the third step, we evaluated the quality of the images for which we performed focus blur detection and black pixel and white pixel estimation. If an image suffers from focus blur or mainly contains black or white pixels, it is eliminated for the analysis. This three-step image augmentation and verification resulted in a balanced and expanded dataset of 23,000 images where each class has 1000 images, as illustrated in Fig. 2

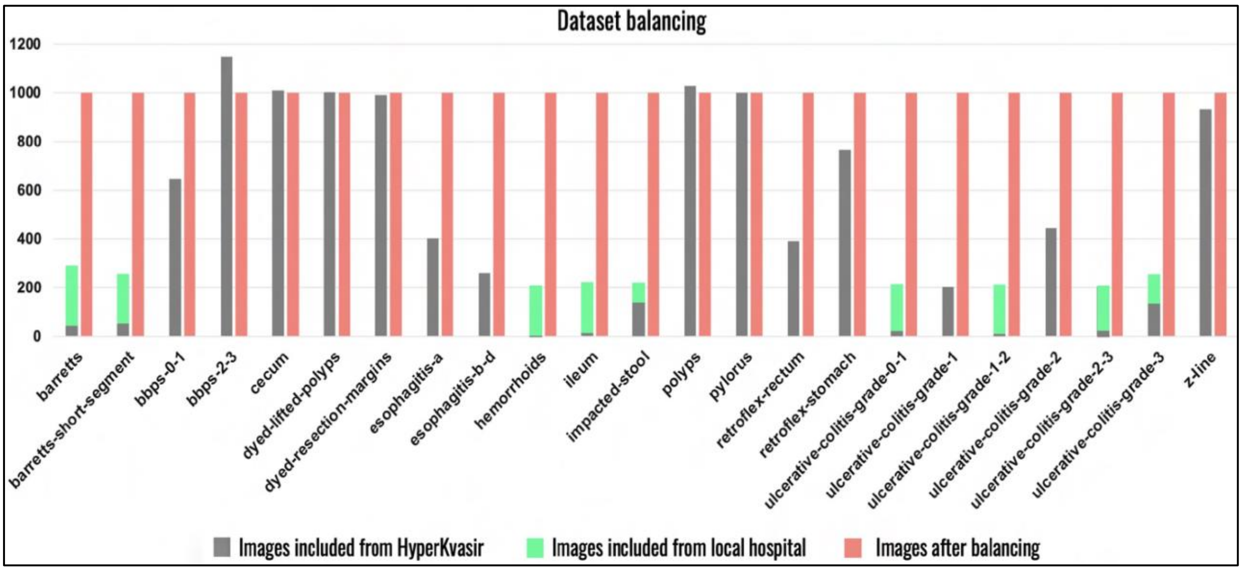

Fig. 2: Adjusting the number of images per class to balance the dataset

The images were then used to train, validate and test machine learning models to find the most suitable network for the proposed system. 13800 (60%), 4600 (20) and 2300 (10%) of the 23,000 images were used for training, validation and testing of the machine learning networks, respectively. Each set has the same number of images for each class. Additionally, we prepared a separate test set, which had 2272 images for the 23 classes, collected from the local hospital. The number of images was different for the classes in this test dataset. Endoscopists utilized these for manual examination. The best DeiT, ViT and CNN-based networks were evaluated using this test set and the results were compared to the results of the manual examination. This is a heterogeneous test set, with images captured using a different endoscopy machine in the local hospital. However, as shown in green in Fig. 2, some images captured in this hospital were included in the training dataset, marginally compromising the heterogeneity of this test set. Both test sets were unseen while the models were being trained and validated.

Image Quality Assessment

Image quality is crucial for training machine learning models because it affects their stability, robustness, practical applicability and generalization capability. High-resolution images are often a fundamental prerequisite for achieving good accuracy of machine learning models in computer vision and related fields, especially for medical image analysis (Hossain *et al*., 2018). The major issues related to the quality of endoscopy images are focus problems, noise, brightness and color (Nishitha *et al*., 2022). Accurate diagnosis and treatment would be compromised in images impacted by these problems. Black and white areas of endoscopic images are additional problems associated with training machine learning models. Machine learning algorithms are confused by dark, black and excessively sharp or mostly white parts that primarily belong to the background. Therefore, the proposed method was designed to eliminate images that contained mostly black or white areas. The proposed system's method for evaluating image quality begins with identifying black or white regions. In a grayscale image, a pixel with an intensity value less than 50 is regarded as black. In contrast, a pixel is considered white if its intensity exceeds 200. An image is rejected for further examination if it contains more than 50% black or white pixels. Subsequently, the images' focus issues and noise are identified. To identify focus error and noise, the proposed method utilized the reference-less image quality evaluation method proposed by Shakhawat *et al*. (2020).

This method incorporated the subjective evaluation of medical practitioners with the objective evaluation to justify their image quality assessment method, which is significant to ensure the practical usability of the method for other medical imaging-related applications. Moreover, this method is reference less and suitable for our work, as finding an ideal endoscopy image is challenging. We utilized this method and estimated the width of the edges as the difference between its local maxima and local minima for the endoscopy images. The sharp edges had a low difference value compared to the blurred edges. Then, the average edge width was calculated for the images. An image was considered blurry or out of focus if its average edge width exceeded five. We considered pixel noise if its intensity value is random and independent from its neighboring pixels for noise detection. Firstly, we produced a blurry version of the original images by applying a Gaussian blur filter to the original images. Then, the blurred version of the image is subtracted from the original one. After that, the minimum difference of pixel values in a 3×3 window was calculated for all the pixels of the resultant images. The minimum difference value is high for a noise pixel with an independent neighborhood value. Finally, the average minimum difference value of the images was used as the noise indicator. An image with an average minimum difference value higher than ten was eliminated as a noisy image. Finally, an image not affected by black pixel, white pixel, blur, or noise artifacts was utilized for the proposed method and transformed to the sRGB color space from RGB. The color transformation was performed to compensate for the color variation. The image quality assessment approach ensured that only good-quality images were used for landmark detection. Moreover, this is crucial to ensure the method's robustness regardless of the endoscopy machines.

Model Training, Evaluation and Selection

Achieving good accuracy in multi-class classification problems is challenging. Classifying a new instance into one of the many classes is more complex than making the same decision where there are fewer classes (Moral *et al*., 2022). Therefore, in this study, we undertook an exhaustive search to select the best deep-learning network to classify anatomical landmarks and abnormalities from one of the 23 classes. This process involved intensive data curation, model selection, network tuning, demonstration and careful evaluation. This study examined the performance of CNN and transformer-based deep learning models for multi-class landmark detection when trained using a limited dataset. The exemplary CNN-based networks used the VGG16, VGG19, InceptionV3, ResNet101, EfficientNet and DenseNet169 models. For the transformer model, ViT and DeiT-based networks were utilized. In the case of ViT, two different base models were utilized. Fine-tuned networks of these models were developed by optimizing the hyper-parameters and training them using the training dataset. Then, the fine-tuned networks of all models were tested for the same test data and compared in terms of their accuracy, precision, recall, F1-score and Area Under the Curve (AUC).

CNNs are explicitly designed for image processing and found effective in medical image analysis. CNN-based models mainly comprise convolutional and pooling operations layers, followed by fully connected layers. Convolutional layers capture local features and pooling layers help reduce spatial dimensions. On the other hand, transformers were originally developed for natural language processing, utilizing a self-attention mechanism. This self-attention technique was later found highly effective in understanding dependencies between image patches, which is not possible using CNN-based architecture. Consequently, the ViT model outperformed the CNN models for various medical imaging applications. However, the ViT required a significantly large amount of data to train and when trained using a small dataset, it often fails to achieve adequate accuracy. The data requirement of ViT becomes more critical for multi-class problems, especially when the number of classes is as large as 23. In the absence of adequate data, the performance of ViT falls significantly. With the exception of an additional distillation token in the input token part, DeiT is a more contemporary transformer model with the same architecture as ViT. Unlike ViT, which requires a large dataset for training, DeiT is better suited for computer vision applications with limited data. DeiT is more suitable for data-efficient computer vision applications, unlike ViT, which requires large datasets. DeiT achieves data efficiency through a combination of techniques and architectural choices, which include knowledge distillation-based regularization, pre-training with ’noisy student,’ and data augmentation. DeiT employs substantial data augmentation during training, exposing the model to a wide range of data variations to make it more resilient. DeiT is prepared using “noisy student," which learns from a wide range of noisy or incorrectly labeled images. This allows the model to capture general features and representations.

The deep learning models were fine-tuned for different combinations of hyperparameters by varying the epoch, batch size, optimizer, loss function, learning rate and dropout values, as shown in Table 2. We utilized the pre-trained weights of these models for fine-tuning using our dataset. The models were pre-trained using the image net dataset (Deng *et al*., 2009). In total, 3,888 networks from six CNN and three transformer models were developed and fine-tuned using 18,400 images of 23 classes, among which 13,800 images were used for training and 4,600 for validation. After that, the networks were tested using 2,300 unseen test images. The best network of these models was compared to select the best network for the proposed system. The comparative results of the best networks of the CNN and transformer-based models are shown in Figs. 4-8 and Table 5.

The DeiT-based network outperformed the other networks in these evaluations and was consequently selected for the proposed system. Furthermore, a separate test set of 2,273 images was prepared, which were examined manually by three experts individually to label the class of each image. Then, these images were also classified using the best DeiT, ViT and CNN-based network separately. The selected DeiT-based network was found to be more accurate and stable when compared to the subjective examination. Figure 3 shows the architecture of the proposed DeiT-based method for landmark and abnormality classification.

This architecture consists of multiple linear layers with batch normalization, ReLU activation and dropout regularization. These layers sequentially extract features and introduce non-linearity. The final linear layer produces the output with a size of 23, representing the classes. The output of the teacher-student network first passes through a linear layer, which applies a linear transformation to produce a tensor of 2048 in size. The result of the linear layer is then normalized using BatchNorm. Non-linearity is introduced by applying ReLU, which sets negative values to zero while preserving positive values.

Table 2: List of hyperparameters and their values explored to fine-tune the CNN and transformer networks

Criteria	Search space
Models	[VGG16, VGG19, inceptionv3, ResNet101, EfficientNet, DenseNet121, ViT-B/16, ViT-B/32, DeiT]
Epochs	[50, 75, 100]
Batch sizes	[16, 32, 64]
Optimizers	[SGD, Adam, AdamW, RMSProp]
Loss functions	[Categorical cross entropy]
Learning rates	[0.01, 0.001, 0.0001]
Dropouts	[0.5, 0.6, 0.7, 0.8]

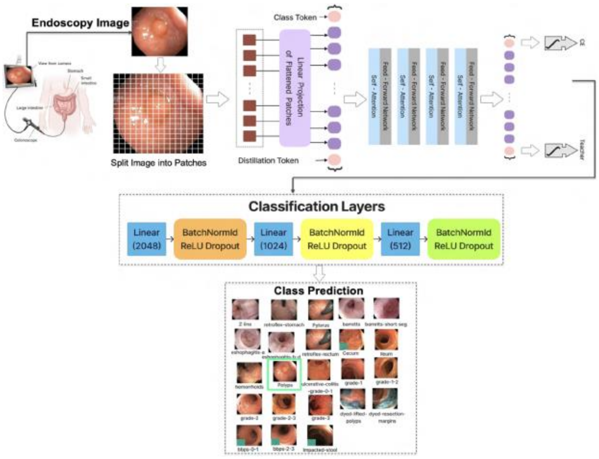

Fig. 3: Architecture for proposed DeiT-based network for landmark and abnormality classification

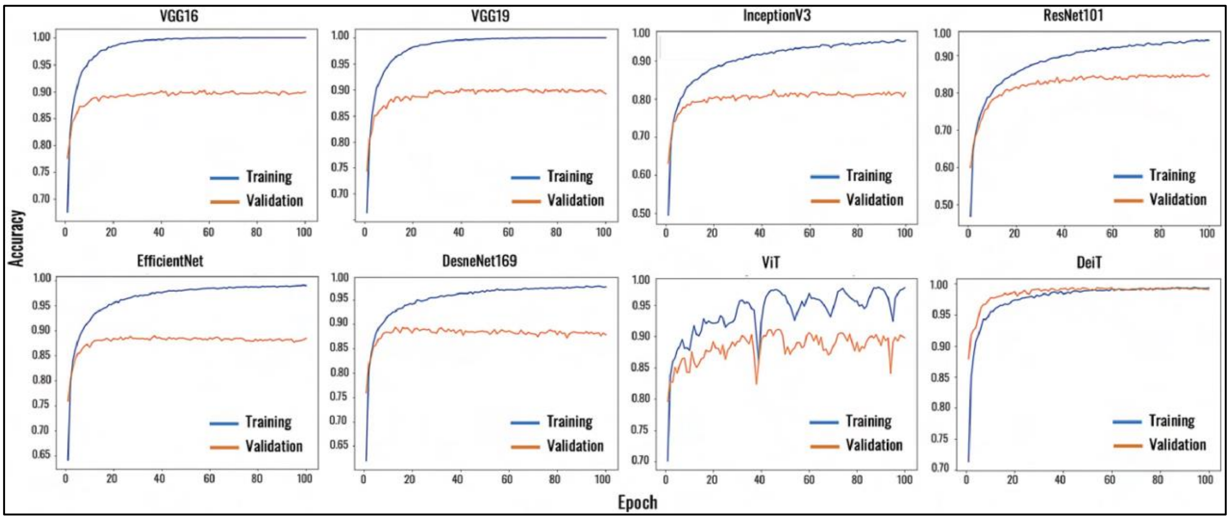

Fig. 4. Training and validation accuracy curves of the models

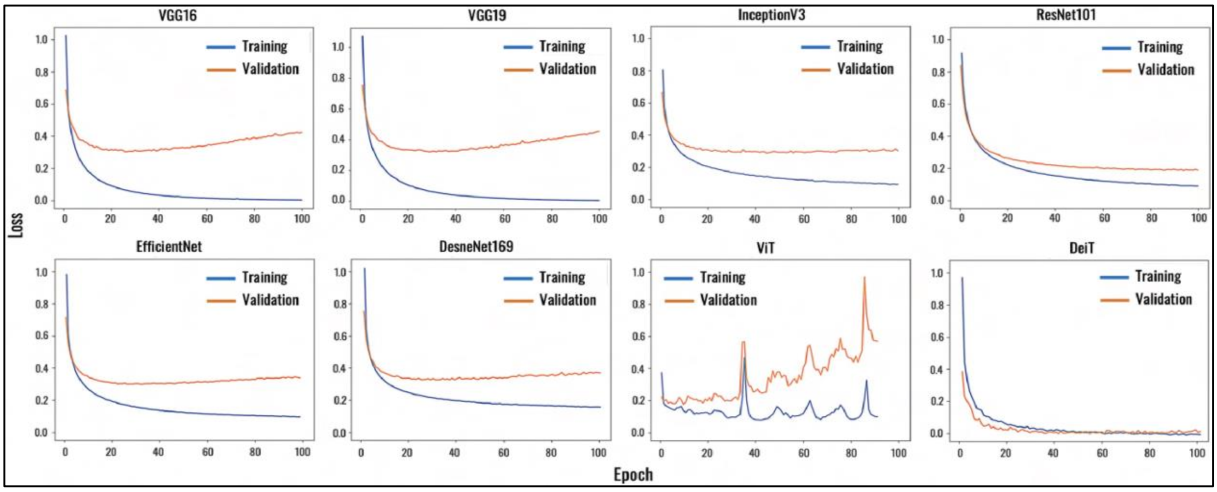

Fig. 5: Training and validation loss curves of the models

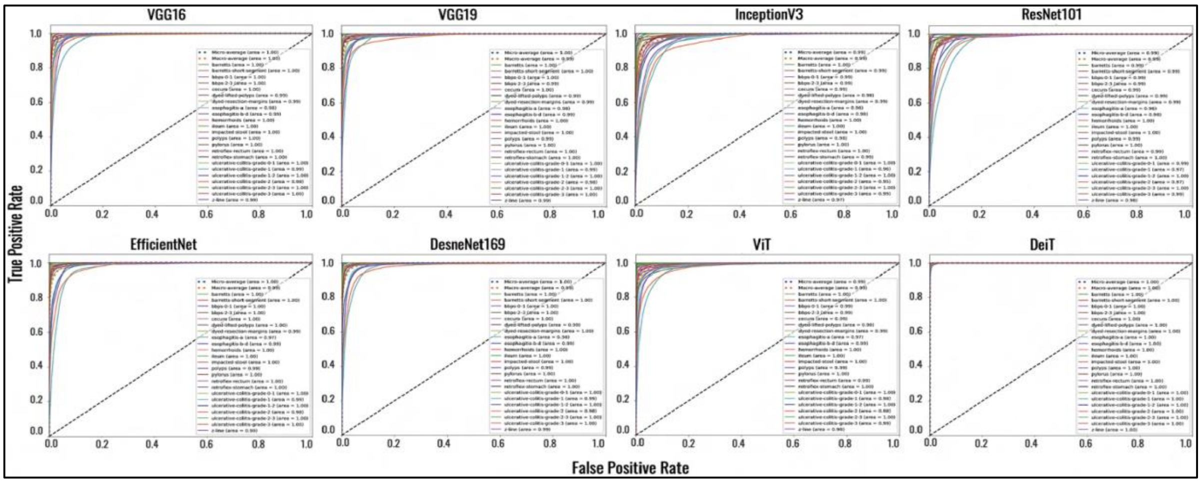

Fig. 6: ROC of machine learning models

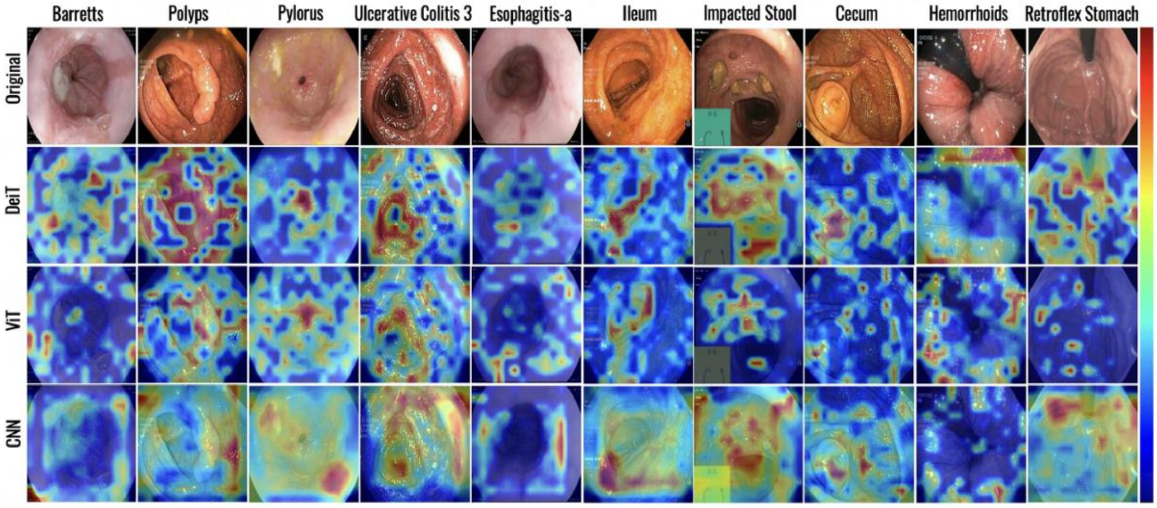

Fig. 7: Grad-Cam representation of best fine-tuned

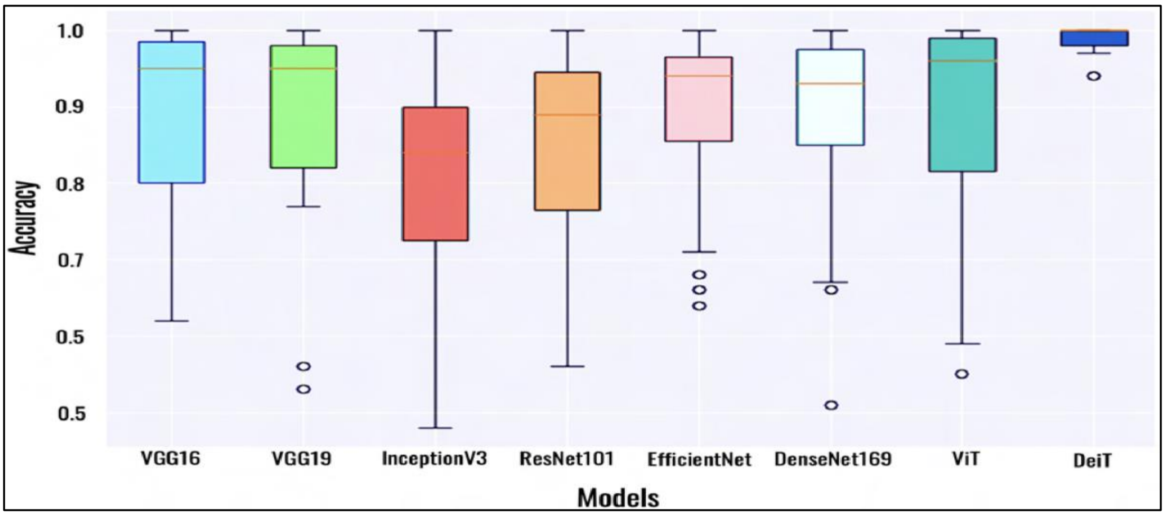

Fig. 8: Comparison of machine learning models using boxplot

Dropout is utilized after ReLU to prevent overfitting by randomly deactivating a percentage of input units during training. The output of the first block is subsequently processed by the second linear layer, which has an output size of 1024, following the same sequence of operations: BatchNorm1d, ReLU and dropout. This process is repeated for the third block, where a linear layer outputs a size of 512. Finally, the output of the third block is fed into the final linear layer, generating a final output of size 23 representing the network's prediction for each class.

Results

Anatomical Landmark and Abnormality Classification

This study fine-tuned and examined the performances of popular CNN-based models, self-attention-based vision transformer model, ViT (Dosovitskiy *et al*., 2020) and recently proposed data efficient transformer model, DeiT (Touvron *et al*., 2021) for classifying anatomical landmarks and abnormalities of GI tract from endoscopy images. For CNN, VGG16 (Simonyan and Zisserman, 2015), VGG19 (Simonyan and Zisserman, 2015), inceptionv3 (Szegedy *et al*., 2015), ResNet101 (He *et al*., 2016), EfficientNet (Tan and Le, 2019) and DenseNet169 (Huang *et al*., 2017) models were used considering their efficiency for medical image analysis. The models were customized by varying the values of hyperparameters, as listed in Table 2 and then trained using 13800 images and validated using 4600 images. Figures 4-5 show the accuracy and loss curves for the best networks of each model for training and validation. Further, these networks were tested on 2300 unseen images to derive the average accuracy, precision, recall, F1-score and Area Under the

Curve (AUC) for 23 classes, as listed in Table 3. The DeiT-based network achieved accuracy, precision, recall and F1-score over 99%, outperforming the ViT and CNN-based networks with significant differences. The AUC was 100% in the Receiver Operating Characteristic (ROC) plot for the DeiT, indicating its strength in distinguishing between the different classes. Figure 6 shows the ROC curves for the networks. The proposed DeiT-based method was also compared with the endoscopist's manual evaluation. For this purpose, a separate test set of 2272 images of 23 classes was prepared. Then, three endoscopists manually examined the images and compared them to the proposed DeiT-based method, shown in Table 1. The proposed method achieved 99% accuracy, outperforming the experts. The DeiT-based method was also compared to the CNN and ViT for the same test set, shown in Table 5. These results explain that the DeiT-based network yielded the highest accuracy and lowest loss; thus, the DeiT-based network was selected for the proposed method.

Table 3: Evaluation of transformer and CNN-based models using test dataset

Networks	Average accuracy	Average precision	Average recall	Average F1-score	Average AUC
VGG16	89.57	89.50	89.56	89.56	99.60
VGG19	88.96	88.65	88.90	88.66	99.52
InceptionV3	80.04	80.25	80.00	79.97	98.73
ResNet101	84.26	84.57	84.25	84.25	98.91
EfficientNet	88.91	88.99	88.91	88.86	99.40
DenseNet169	88.17	88.09	88.17	87.91	99.60
ViT	88.19	88.30	88.21	88.52	99.41
DeiT	99.65	99.34	99.34	99.35	100.00

Table 4: Comparison between existing automated methods and the proposed method

Criteria	Proposed method	Borgli *et al.* (2020) (state of the art)	Che *et al.* (2021)	Tran *et al.* (2021)
Dataset	Hyper kvasir + Local Hospital data (23 landmarks from entire GI tract)	Hyper kvasir (23 landmarks from entire GI tract)	Local hospital data (3 landmarks from lower GI tract)	Local hospital data (10 landmarks from upper GI tract)
Accuracy	99.650	91.00	92.03	97.43
Precision	99.340	91.00	92.34	97.43
Recall	99.340	91.00	91.01	97.43
F1-score	99.350	91.00	91.57	-
AUC	100.000	-	-	-
Phi-coefficient	99.700	90.20	-	-
Time (second/image)	0.045±0.005	-	-	-

Table 5: Class-wise accuracy of CNN, ViT and DeiT networks for the selected test images

Anatomical landmarks and abnormalities	ViT accuracy %	CNN accuracy %	DeiT accuracy %
Barrett	82/86 (96)	86/86 (100)	86/86 (100)
Barrett's short segment	95/99 (96)	99/99 (100)	99/99 (100)
BBPS 0-1	96/100 (96)	98/100 (98)	100/100 (100)
BBPS 2-3	73/77 (95)	74/77 (96)	76/77 (98)
Cecum	97/116 (84)	100/116 (86)	115/116 (99)
Polyps	85/107 (80)	87/107 (81)	106/107 (99)
Dyed lifted polyps	73/97 (76)	83/97 (85)	97/97 (100)
Dyed resection margin	73/93 (78)	77/93 (82)	91/93 (97)
Esophagitis-a	70/99 (70)	79/99 (79)	99/99 (100)
Esophagitis b-d	93/107 (86)	87/107 (81)	106/107 (99)
Hemorrhoids	102/102 (100)	102/102 (100)	102/102 (100)
Ileum	101/101 (100)	101/101 (100)	101/101 (100)
Impacted stool	95/96 (98)	95/96 (99)	96/96 (100)
Pylorus	101/102 (99)	100/102 (98)	102/102 (100)
Retroflex rectum	92/100 (92)	96/100 (96)	100/100 (100)
Retroflex stomach	110/112 (98)	110/112 (98)	112/112 (100)
Ulcerative colitis 0-1	80/81 (99)	80/81 (99)	81/81 (100)
Ulcerative colitis 1-2	107/107 (100)	107/107 (100)	107/107 (100)
Ulcerative colitis 2-3	110/110 (100)	109/110 (99)	110/110 (100)
Ulcerative colitis 1	73/97 (75)	70/97 (72)	97/97 (100)
Ulcerative colitis 2	49/78 (62)	43/78 (55)	78/78 (100)
Ulcerative colitis 3	98/104 (94)	94/104 (90)	104/104 (100)
Z-line	77/101 (76)	78/101 (77)	101/101 (100)
Overall	2032/2272 (89)	2055/2272 (90)	2266/2272 (99)

Finally, we compared the results of the proposed method derived from the test data with the relevant existing automated endoscopy image classification methods, as shown in Table 4. The proposed method classified the maximum classes and tested on the most diverse and heterogeneous dataset collected from two labs and captured using two separate machines. Despite that, the proposed method outperformed the existing methods in all criteria. Furthermore, we applied gradient-weighted class activation mapping (Selvaraju *et al*., 2017) (grad-CAM) on the images to interpret the network's intelligence, as shown in Fig. 7.

Analysis of CNN and Transformer for Endoscopy Examination

The ViT (Dosovitskiy *et al*., 2020) model's self-attention technique was recently found to be highly effective in understanding dependencies between regions of pathological images and outperformed the conventional CNN models. However, training the ViT and CNN models to achieve an accurate and stable predictive network requires large data, particularly when the number of classes to predict is high. Moreover, these models lose accuracy and stability as the number of classes increases. Unlike CNN and ViT, the DeiT is claimed to be data efficient by the author Touvron *et al*. (2021). In this study, we trained the CNN, ViT and DeiT using the same data and compared their performances. Preparing a large dataset is one of the main challenges for training AI for medical imaging analysis. This study examined how well the models performed when trained using a smaller dataset, with an average of 320 images per class before augmentation. Considering that there are up to 23 classes, this is remarkably low.

The ViT failed to achieve stability in 100 epochs, as shown in Fig. 4. Although the accuracies of CNN-based networks were comparable to ViT, they were primarily over-fitted, as indicated by their loss curves in Fig. 5. Table 3 shows ViT was marginally outperformed by the VGG architecture-based networks VGG16 and VGG19, indicating that ViT is confined to applications with large amounts of data. In the comparison using a separate test set of 2272 images, shown in Table 5, the CNN-based network (90.44%) marginally outperformed the ViT (89.43%) while the DeiT achieved the highest accuracy of 99.73%. Further, we plotted the test accuracies for the best network of each model using a boxplot, shown in Fig. 8. The DeiT had the highest median value, which is the mid-point of the data, indicated by an orange horizontal line. The boxplots also showed that the DeiT has a narrow accuracy spread over the images, demonstrating that DeiT is more consistent than other models. The comparative results of CNN, ViT and DeiT demonstrated that, when trained on a comparatively smaller dataset for 23 classes, the DeiT-based method yielded the best results for classifying landmarks and abnormalities. The tendency of CNN-based networks to overfit may be attributed to the fact that the original image counts in the training dataset were much lower for some classes, such as those related to Ulcerative Colitis. The CNN and ViT-based networks also resulted in most misclassification for these classes. However, DeiT was not affected by such a problem.

We also plotted the grad-CAM for the models, which helped to visualize how the networks are using the image to predict its class by using a layer's gradient to get the layer's attention. The grad-CAM utilized the last convolutional layer for the CNN. The gradient of the last attention layer was utilized for the ViT and DeiT with no convolutional block. Fig. 7 shows the grad-CAM representation for ten major classes, visualizing how each model utilized the image to determine its class. Regions with higher network priorities for predictions are highlighted in red color. It indicates that the DeiT outperformed ViT and CNN in locating the image's critical areas for the prediction. The transformer's comprehension of pixel dependencies can be explained by the fact that CNN incorporates more background or irrelevant pixels than ViT and DeiT. DeiT was found to be more aggressive and precise in prioritizing the regions of the image for predicting its class. The DeiT failed to link the image information like human experts for hemorrhoids and esophagitis. Nevertheless, it succeeded in identifying the classes accurately despite the fact. The DeiT model did not fully exploit esophageal ulcer pixels in the case of esophagitis images. This also applies to CNN and ViT models.

Based on the findings above, this study concludes that the DeiT-based network is more suited for the proposed system since it is more appropriate than CNN and ViT for data-efficient image classification problems, which is our case in this research.

Feasibility of Proposed Method for Clinical Use

We assessed the feasibility of the proposed method for clinical use in terms of its accuracy, speed and diverseness. The accuracy of the proposed was compared with both objective scores of existing relevant automated methods and subjective evaluation scores of experts. Table 4 shows the comparison with the automated methods. Table 1 shows the comparison with the experts' results. For this subjective evaluation, three experts manually examined the 2,272 endoscopy images individually on a computer screen. These 2,272 images were clinically diagnosed previously to establish the ground truth. In this experiment, the endoscopy images were provided randomly to the experts through a website to assign a label from one of the 23 classes. The website had the facility to zoom and pan the images. It also allowed the experts to provide comments on the images. The proposed method was tested for the same test set. Finally, the results were compared in Table 1. The proposed DeiT method classified 2266 (99.73%) images correctly out of 2272 images, where experts A, B and C identified 2084 (91.72%), 2133 (93.88%) and 2175 (95.73%) images, respectively. It has been observed that significant misclassification occurred for the different grades of ulcerative colitis. A few cecum images were misidentified as ileum and BBPS 0-1 images as BBPS 2-3 by the experts. However, such a misidentification pattern was not observed for the proposed method. We also estimated the Phi coefficient, known as the Matthew correlation coefficient, for the proposed method and experts by comparing them with the clinical ground truth. The Phi coefficient is a more reliable statistical measure that produces a high score only if the method performs well in all four confusion matrix indexes: True positives, false negatives, true negatives and false positives (Chicco and Jurman, 2020). The Phi coefficients for experts A, B and C were 0.913, 0.936 and 0.950, respectively. The coefficient was 0.997 for the proposed method, indicating higher reliability than the existing manual examination.

Further, the time complexity of the proposed method was estimated to determine its suitability for practical use.

The proposed method took an average of 100±20 sec to classify 2300 images. A personal computer with an 11th-generation Intel Core i5 2.40 GHz processor and 8 GB RAM was used to estimate the time. The experts' manual examination time was far longer than the proposed method and the state-of-the-art did not disclose its time. The proposed method can classify 23 images on average in a second. It follows that the proposed method is accurate, rapid and diverse enough for autonomous endoscopy examination in clinical practice.

Discussion

Endoscopy is a challenging, invasive procedure that depends on various factors, including the type of endoscopy, the specific organ being examined, the patient's condition and the endoscopist's experience and skill. The endoscopist has to continuously identify different landmarks and abnormalities from the video to make appropriate navigation, diagnosis and surgery decisions in real time. An automated method for detecting and classifying landmarks and abnormalities can significantly ease the task for endoscopists and improve patient experience. In this study, we present an AI-assisted landmark and abnormality detection method for automated endoscopy examination to assist endoscopists in endoscopy. The proposed system is designed to identify 23 landmarks and abnormalities, the highest number of classes any study considered. This method achieved 99% accuracy when tested on two heterogeneous datasets, which is 8% higher than the current state of the art. Additionally, this method was compared with experts' manual examination results in which it outperformed the experts. Therefore, it can be concluded that the proposed method enables more accurate and reliable endoscopy examination with less effort. The proposed system can examine 2300 images in less than 120 sec, which ensures its rapidness. The high accuracy and rapid detection time suggest that the existing manual examination can be replaced with the proposed automated method. Another essential aspect of this study is that the proposed method not only improves the accuracy of the diagnosis but also enables the endoscope to be maneuvered more effectively, improving patient comfort and lowering the risk of injury. This study also evaluated the performance of CNN and transformer models using limited images. The results revealed that the DeiT, a knowledge distillation-based transformer, is more suitable for computer vision applications with relatively small datasets than CNN and other transformers, which is another important finding of this study.

There are some challenges and limitations of this study. One of the significant challenges was handling the image artifacts. Major artifacts found in the endoscopy image were cyan marks, text marks and white and black pixels. Artifacts mislead AI models and result in wrong diagnoses.

Therefore, the proposed system was designed to detect the white and black pixels to eliminate images highly affected by these artifacts. Black backgrounds are the most troublesome artifact for training the AI model as they are similar to blood regions or hemorrhoids. The proposed system did not include cyan marks detection, a limitation to be fixed soon. Another challenge of this study was that certain classes had significantly fewer images than others, making the dataset unbalanced. We augmented the images to balance the dataset and then trained the models using the balanced dataset. However, an augmented dataset generated from minimal original images risks the generalization ability of the AI models. Therefore, we tested the proposed method on a separate dataset to ensure its generalization ability and robustness, in which it achieved 99% accuracy. Nonetheless, a small number of images captured at the same hospital were included in the training dataset. This slightly impacted the robustness and heterogeneity since the test and train dataset had different images prepared at the same hospital for a few classes. However, given the quantity of images, this can be disregarded. The proposed method does not correlate the image pixels like experts do for some landmarks. As demonstrated in Fig. 7, the esophageal ulcer shown in the image, for instance, lacks a corresponding red heat map indicating esophagitis. This is a drawback of the suggested approach. However, the accuracy of the proposed method in identifying esophagitis is adequate. In some situations, it is not feasible to map the precise relationships between how AI-based techniques and human specialists analyze visual data.

The results of this study demonstrate that the proposed method will improve diagnosis accuracy, enable better navigation and enhance patient comfort. The proposed method can be extended for the autonomous navigation of endoscopy.

Conclusion

In this study, we proposed an image transformer-based method for detecting anatomical landmarks, abnormalities and pathological findings to guide endoscopists for better movement of the endoscope and precise diagnosis. The effectiveness of the proposed method was validated using objective measures and subjective evaluation by expert endoscopists. This system will facilitate autonomous endoscopy examination, reducing time and labor and improving patient satisfaction.

Acknowledgment

The authors thank Dr. Md. Aulad Hossain, Dr. Zakia Zahan, Dr. Anwer Hussain and Dr. Sahria Bakar for their contribution in preparing the dataset and evaluating the proposed system.

Funding Information

This research is funded by the RIoT center, independent university, Bangladesh, Dhaka, Bangladesh (http://www.riot.iub.ac.bd/).

Author’s Contributions

All authors equally contributed to this study.

Ethics

Deidentified human endoscopic images were used in this study to ensure that the principle of self-determination is not violated. We have received approval (IRB 01-26-2022) from the Institutional Review Board (IRB) committee of independent university, Bangladesh. The IRB committee waived the need for informed consent for this study since the data was anonymous and it was not feasible to get each patient’s consent verbally or in writing.

Data Availability

The hyper kvasir dataset utilized for this study is available in the hyper kvasir. Hyper kvasir is an open-access dataset and licensed under a Creative Commons attribution 4.0 international (CC BY 4.0). The endoscopic images collected from the local hospital are not publicly available now.

References

Aabakken, L., Barkun, A. N., Cotton, P. B., Fedorov, E., Fujino, M. A., Ivanova, E., Kudo, S., Kuznetzov, K., de Lange, T., Matsuda, K., Moine, O., Rembacken, B., Rey, J., Romagnuolo, J., Rösch, T., Sawhney, M., Yao, K., & Waye, J. D. (2014). Standardized endoscopic reporting. *Journal of Gastroenterology and Hepatology*, *29*(2), 234-240. https://doi.org/10.1111/jgh.12489

Aliyi, S., Dese, K., & Raj, H. (2023). Detection of gastrointestinal tract disorders using deep learning methods from colonoscopy images and videos. *Scientific African*, *20*, e01628. https://doi.org/10.1016/j.sciaf.2023.e01628

Ayyoubi Nezhad, S., Khatibi, T., & Sohrabi, M. (2022). Proposing Novel Data Analytics Method for Anatomical Landmark Identification from Endoscopic Video Frames. *Journal of Healthcare Engineering*, *2022*, 8151177. https://doi.org/10.1155/2022/8151177

Bour, A., Castillo-Olea, C., Garcia-Zapirain, B., & Zahia, S. (2019). Automatic colon polyp classification using Convolutional Neural Network: A Case Study at Basque Country. *2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 1-5. https://doi.org/10.1109/isspit47144.2019.9001816

Borgli, H., Thambawita, V., Smedsrud, P. H., Hicks, S., Jha, D., Eskeland, S. L., Randel, K. R., Pogorelov, K., Lux, M., Nguyen, D. T. D., Johansen, D., Griwodz, C., Stensland, H. K., Garcia-Ceja, E., Schmidt, P. T., Hammer, H. L., Riegler, M. A., Halvorsen, P., & de Lange, T. (2020). HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data*, *7*, 283. https://doi.org/10.1038/s41597-020-00622-y

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*, 6. https://doi.org/10.1186/s12864-019-6413-7

Che, K., Ye, C., Yao, Y., Ma, N., Zhang, R., Wang, J., & Meng, M. Q. H. (2021). Deep learning-based biological anatomical landmark detection in colonoscopy videos. *ArXiv*, 9 pages. https://doi.org/10.48550/arXiv.2108.02948

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fe-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248-255. https://doi.org/10.1109/cvprw.2009.5206848

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv*, 2010.11929. https://doi.org/10.48550/arXiv.2010.11929

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D. M., Forman, D., & Bray, F. (2015). Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *International Journal of Cancer*, *136*(5), E359-E386. https://doi.org/10.1002/ijc.29210

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 770-778. https://doi.org/10.1109/cvpr.2016.90

Hirasawa, T., Aoyama, K., Tanimoto, T., Ishihara, S., Shichijo, S., Ozawa, T., Ohnishi, T., Fujishiro, M., Matsuo, K., Fujisaki, J., & Tada, T. (2018). Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images. *Gastric Cancer*, *21*, 653-660. https://doi.org/10.1007/s10120-018-0793-2

Hossain, M. S., Nakamura, T., Kimura, F., Yagi, Y., & Yamaguchi, M. (2018). Practical image quality evaluation for whole slide imaging scanner. *Biomedical Imaging and Sensing Conference*. Biomedical Imaging and Sensing Conference, Yokohama, Japan. https://doi.org/10.1117/12.2316764

Hossain, M. S., Rahman, Md. M., Syeed, M. M., Uddin, M. F., Hasan, M., Hossain, Md. A., Ksibi, A., Jamjoom, M. M., Ullah, Z., & Samad, M. A. (2023). DeepPoly: Deep Learning-Based Polyps Segmentation and Classification for Autonomous Colonoscopy Examination. *IEEE Access*, *11*, 95889-95902. https://doi.org/10.1109/access.2023.3310541

Hossain, M. S., Syeed, M. M. M., Fatema, K., & Uddin, M. F. (2022). The Perception of Health Professionals in Bangladesh toward the Digitalization of the Health Sector. *International Journal of Environmental Research and Public Health*, *19*(20), 13695. https://doi.org/10.3390/ijerph192013695

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261-2269. https://doi.org/10.1109/cvpr.2017.243

Iwagami, H., Ishihara, R., Aoyama, K., Fukuda, H., Shimamoto, Y., Kono, M., Nakahira, H., Matsuura, N., Shichijo, S., Kanesaka, T., Kanzaki, H., Ishii, T., Nakatani, Y., & Tada, T. (2021). Artificial intelligence for the detection of esophageal and esophagogastric junctional adenocarcinoma. *Journal of Gastroenterology and Hepatology*, *36*(1), 131-136. https://doi.org/10.1111/jgh.15136

Kaminski, M., F., Regula, J., Kraszewska, E., Polkowski, M., Wojciechowska, U., Didkowska, J., Zwierko, M., Rupinski, M., Nowacki, M. & Butruk, E. (2010). Quality indicators for colonoscopy and the risk of interval cancer. *New England Journal of Medicine*. *362*(19), 1795-1803. https://doi.org/10.1056/NEJMoa0907667

Luo, H., Xu, G., Li, C., He, L., Luo, L., Wang, Z., Jing, B., Deng, Y., Jin, Y., Li, Y., Li, B., Tan, W., He, C., Seeruttun, S. R., Wu, Q., Huang, J., Huang, D., Chen, B., Lin, S., … Xu, R. (2019). Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: A multicentre, case-control, diagnostic study. *The Lancet Oncology*, *20*(12), 1645-1654. https://doi.org/10.1016/s1470-2045(19)30637-0

Misawa, M., Kudo, S., Mori, Y., Hotta, K., Ohtsuka, K., Matsuda, T., Saito, S., Kudo, T., Baba, T., Ishida, F., Itoh, H., Oda, M., & Mori, K. (2021). Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). *Gastrointestinal Endoscopy*, *93*(4), 960-967. https://doi.org/10.1016/j.gie.2020.07.060

Moral, P. D., Nowaczyk, S., & Pashami, S. (2022). Why Is Multiclass Classification Hard? *IEEE Access*, *10*, 80448-80462. https://doi.org/10.1109/access.2022.3192514

Nishitha, R., Amalan, S., Sharma, S., Preejith, S. P., & Sivaprakasam, M. (2022). Image Quality Assessment for Interdependent Image Parameters Using a Score-Based Technique for Endoscopy Applications. *2022 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, 1-6. https://doi.org/10.1109/memea54994.2022.9856448

Ozawa, T., Ishihara, S., Fujishiro, M., Kumagai, Y., Shichijo, S., & Tada, T. (2020). Automated endoscopic detection and classification of colorectal polyps using convolutional neural networks. *Therapeutic Advances in Gastroenterology*, *13*, 175628482091065. https://doi.org/10.1177/1756284820910659

Pogorelov, K., Randel, K. R., Griwodz, C., Eskeland, S. L., de Lange, T., Johansen, D., Spampinato, C., Dang-Nguyen, D.-T., Lux, M., Schmidt, P. T., Riegler, M., & Halvorsen, P. (2017). KVASIR: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection. *Proceedings of the 8th ACM on Multimedia Systems Conference*, 164-169. https://doi.org/10.1145/3083187.3083212

Suzuki, H., Yoshitaka, T., Yoshio, T., & Tada, T. (2021). Artificial intelligence for cancer detection of the upper gastrointestinal tract. *Digestive Endoscopy*, *33*(2), 254-262. https://doi.org/10.1111/den.13897

Shakhawat, H. M., Nakamura, T., Kimura, F., Yagi, Y., & Yamaguchi, M. (2020). [Paper] Automatic Quality Evaluation of Whole Slide Images for the Practical Use of Whole Slide Imaging Scanner. *ITE Transactions on Media Technology and Applications*, *8*(4), 252-268. https://doi.org/10.3169/mta.8.252

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *ArXiv*, 1409.1556. https://doi.org/10.48550/arXiv.1409.1556

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *2017 IEEE International Conference on Computer Vision (ICCV)*, 618-626. https://doi.org/10.1109/iccv.2017.74

Szegedy, C., Wei Liu, Yangqing Jia, Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1-9. https://doi.org/10.1109/cvpr.2015.7298594

Tomar, N. K., Jha, D., Ali, S., Johansen, H. D., Johansen, D., Riegler, M. A., & Halvorsen, P. (2021). DDANet: Dual Decoder Attention Network for Automatic Polyp Segmentation. In A. Del Bimbo, R. Cucchiara, S. Sclaroff, G. Maria Farinella, T. Mei, M. Bertini, H. Jair Escalante, & R. Vezzani (Eds.), *Pattern Recognition. ICPR International Workshops and Challenges* (Vol. *12668*, pp. 307-314). Springer, Cham. https://doi.org/10.1007/978-3-030-68793-9_23

Tran, T.-H., Nguyen, P.-T., Tran, D.-H., Manh, X.-H., Vu, D.-H., Ho, N.-K., Do, K.-L., Nguyen, V.-T., Nguyen, L.-T., Dao, V.-H., & Vu, H. (2021). Classification of anatomical landmarks from upper gastrointestinal endoscopic images⋆. *2021 8th NAFOSTED Conference on Information and Computer Science (NICS)*, 278-283. https://doi.org/10.1109/nics54270.2021.9701513

Touvron, H., Sablayrolles, A., Douze, M., Cord, M., & Jegou, H. (2021). Grafit: Learning fine-grained image representations with coarse labels. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 854-864. https://doi.org/10.1109/iccv48922.2021.00091

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning, 97,* 6105-6114.